Graduate Texts in Mathematics

Rainer Kress

Numerical Analysis

Springer

Graduate Texts in Mathematics 181

Springer Science+Business Media, LLC

Graduate Texts in Mathematics

continued after index

Rainer Kress

Numerical Analysis

With 11 Illustrations

Springer

Rainer Kress
Institut für Numerische und
Angewandte Mathematik
Universität Göttingen
D-37083 Göttingen
Germany

Mathematics Subject Classification (1991): 65-01

Library of Congress Cataloging-in-Publication Data
Kress, Rainer, 1941-
Numerical analysis / Rainer Kress.
p. cm. – (Graduate texts in mathematics ; 181)
Includes bibliographical references and index.
ISBN 978-1-4612-6833-8 ISBN 978-1-4612-0599-9 (eBook)
DOI 10.1007/978-1-4612-0599-9
1. Numerical analysis. I. Title. II. Series.
QA297.K725 1998
519.4–dc21 97-43748

Printed on acid-free paper.

Production managed by Karina Mikhli; manufacturing supervised by Jacqui Ashri.
Photocomposed copy prepared from the author's TeX file.

9 8 7 6 5 4 3 2 1

ISBN 978-1-4612-6833-8

Preface

No applied mathematician can be properly trained without some basic understanding of numerical methods, i.e., *numerical analysis.* And no scientist and engineer should be using a package program for numerical computations without understanding the program's purpose and its limitations. This book is an attempt to provide some of the required knowledge and understanding. It is written in a spirit that considers numerical analysis not merely as a tool for solving applied problems but also as a challenging and rewarding part of mathematics. The main goal is to provide insight into numerical analysis rather than merely to provide numerical recipes.

The book evolved from the courses on numerical analysis I have taught since 1971 at the University of Göttingen and may be viewed as a successor of an earlier version jointly written with Bruno Brosowski [10] in 1974. It aims at presenting the basic ideas of numerical analysis in a style as concise as possible. Its volume is scaled to a one-year course, i.e., a two-semester course, addressing second-year students at a German university or advanced undergraduate or first-year graduate students at an American university.

In order to make the book accessible not only to mathematicians but also to scientists and engineers, I have planned it to be as self-contained as possible. As prerequisites it requires only a solid foundation in differential and integral calculus and in linear algebra as well as an enthusiasm to see these fundamental and powerful tools in action for solving applied problems. A short presentation of some basic functional analysis is provided in the book to the extent required for a modern presentation of numerical analysis and a deeper understanding of the subject.

An introductory book of a few hundred pages cannot completely cover all classical aspects of numerical analysis and all of the more recent developments. I am willing to admit that the choice of some of the topics in the present volume is biased by my own preferences and that some important subjects are omitted.

I was taught numerical analysis in the mid sixties by my thesis adviser, Professor Erich Martensen, at the Technische Hochschule in Darmstadt. Martensen's perspective on teaching mathematics in general and numerical analysis in particular had a great and long-lasting impact on my own teaching. Therefore, this book is dedicated to Erich Martensen on the occasion of his seventieth birthday.

I would like to thank Thomas Gerlach and Peter Otte for carefully reading the book, for checking the solutions to the problems, and for a number of suggestions for improvements. Special thanks are given to my friend David Colton for reading over the book for correct use of the English language. Part of the book was written while I was on sabbatical leave at the Department of Mathematical Sciences at the University of Delaware and the Department of Mathematics at the University of New South Wales. I gratefully acknowledge the hospitality of these institutions. I also am grateful to Springer-Verlag for being willing to take the economic risk of adding yet another volume to the already huge number of existing introductions to numerical analysis.

Göttingen, September 1997 Rainer Kress

Contents

Glossary of Symbols

Sets and Spaces

$\mathbb{N}$	set of natural numbers
$\mathbb{Z}$	set of integers
$\mathbb{R}$	set of real numbers
$\mathbb{C}$	set of complex numbers
$\|x\|$	absolute value of a real or complex number x
(a, b)	open interval $(a, b) := \{x \in \mathbb{R} : a < x < b\}$
$[a, b]$	closed interval $[a, b] := \{x \in \mathbb{R} : a \leq x \leq b\}$
$\overline{x}$	conjugate of a complex number x
$\mathbb{R}^n$	n-dimensional real Euclidean space
$\mathbb{C}^n$	n-dimensional complex Euclidean space
$C[a, b]$	space of real- or complex-valued continuous functions on the interval $[a, b]$
$C^m[a, b]$	space of m-times continuously differentiable functions
$L^2[a, b]$	space of real- or complex-valued square-integrable functions
$\{a_1, \dots, a_m\}$	set of m elements $a_1, \dots, a_m$
$U \times V$	product $U \times V := \{(x, y) : x \in U,\, y \in V\}$ of two sets U and V
$U \setminus V$	difference set $U \setminus V := \{x \in U : x \notin V\}$ for two sets U and V
$\overline{U}$	closure of a set U
$F : X \to Y$	a mapping with domain X and range in Y

Vectors and Matrices

$x = (x_1, \ldots, x_n)$	row vector in $\mathbb{R}^n$ or $\mathbb{C}^n$ with components $x_1, \ldots, x_n$
$x^T = (x_1, \ldots, x_n)^T$	the transpose of x, i.e., a column vector
$x^* = (\bar{x}_1, \ldots, \bar{x}_n)^T$	the adjoint of x
$A = (a_{jk})$	$m \times n$ matrix with elements a_{jk}
A^T	the transpose of A
A^*	the adjoint of A
$A^\dagger$	the pseudo-inverse of A
A^{-1}	the inverse of an $n \times n$ matrix A
$\det A$	the determinant of an $n \times n$ matrix A
$\mathrm{cond}(A)$	the condition number of an $n \times n$ matrix A
$\rho(A)$	the spectral radius of an $n \times n$ matrix A
I	the $n \times n$ identity matrix
$\mathrm{diag}(a_1, \ldots, a_n)$	diagonal matrix with diagonal elements $a_1, \ldots, a_n$

Norms

$\|\cdot\|$	norm on a linear space
$\|\cdot\|_1$	ℓ_1 norm of a vector, L_1 norm of a function
$\|\cdot\|_2$	ℓ_2 norm of a vector, L_2 norm of a function
$\|\cdot\|_\infty$	maximum norm of a vector or a function
$(\cdot, \cdot)$	scalar product on a linear space

Miscellaneous

$\in$	element inclusion
$\subset$	set inclusion
$\cup, \cap$	union and intersection of sets
$\emptyset$	empty set
$O(m)$	a quantity of order m
$\square$	end of proof

1
Introduction

Numerical analysis is concerned with the development and investigation of *constructive methods* for the numerical solution of mathematical problems. This objective differs from a pure-mathematical approach as illustrated by the following three examples.

By the fundamental theorem of algebra, a polynomial of degree n has n complex zeros. The various proofs of this result, in general, are nonconstructive and give no procedure for the explicit computation of these zeros. Numerical analysis provides constructive methods for the actual computation of the zeros of a polynomial.

The solution of a system of n linear equations for n unknowns can be given explicitly by Cramer's rule. However, Cramer's rule is only of theoretical importance, since for actual computations it is completely useless for linear systems with more than three unknowns. An important task in numerical analysis consists in describing and developing more practical methods for the solution of systems of linear equations.

By the Picard–Lindelöf theorem, the initial value problem for an ordinary differential equation has a unique solution (under appropriate regularity assumptions). Despite the fact that the existence proof in the Picard–Lindelöf theorem actually is constructive through the use of successive iterations, in applied mathematics there is need for more effective procedures to numerically solve the initial value problem.

In general, we may say that for the basic problems in numerical analysis existence and uniqueness of a solution are guaranteed through the results of pure mathematics. The main topic of numerical analysis is to provide efficient numerical methods for the actual computation of the solution. In

some cases these numerical methods are actually based on constructive existence proofs.

By a constructive method we understand a procedure that for any prescribed accuracy determines an approximate solution by a finite number of computational steps. In general, the number of computational steps of course will depend on the required accuracy. Only very few methods will terminate with the exact solution after finitely many computational steps as, for example, Gaussian elimination for solving a system of linear equations. In most cases, the numerical methods will only yield approximations to the exact solution. As a typical example, the numerical evaluation of a definite integral by the trapezoidal rule will, in general, provide only an approximate value for the integral. In this context two main questions arise, namely the question of estimating the error between the exact and the approximate solution and the question of numerical stability.

A numerical method is useful only if it is possible to decide on the accuracy of the approximate solution, i.e., if reliable estimates on the difference between the exact and approximate solution can be given. Therefore, besides the development and design of numerical schemes, a substantial part of numerical analysis is concerned with the investigation and estimation of the errors occurring in these schemes. Here one has to discriminate between the approximation errors, i.e., the errors that arise through replacing the original problem by an approximate problem, and the roundoff errors, i.e., the errors that occur through the fact that in the actual computation, in general, real numbers are replaced by floating-point decimal numbers with a fixed number of digits.

As far as stability is concerned, one has to distinguish between properly and improperly posed problems. A problem is called properly posed or well-posed if the solution depends continuously on the data, i.e., if small changes in the data cause only small changes in the solution. Otherwise, the problem is called improperly posed or ill-posed. Numerical approximations never can circumvent the improper posedness of a problem. However, it is desirable to control the effects of the ill-posed nature of a problem by an adequate choice of the numerical method. On the other hand, for properly posed problems efforts have to be made not to destroy the well-posedness by a poorly designed numerical approximation.

To the author's taste, the topic of stability and properly posedness is more challenging from a mathematical perspective than the rather uninspiring topic of roundoff errors. Therefore, in this book emphasis is given to ill-posedness and the related issue of ill-conditioning, whereas the discussion of roundoff errors is given only cursory attention.

The basic problems of numerical analysis are as old as mathematics itself, and for a number of problems there exist classical approaches such as Newton's method for the solution of nonlinear equations, Gaussian elimination for the solution of systems of linear equations, Gauss–Seidel and Jacobi iterations for linear systems, Lagrange interpolation for the ap-

proximation of arbitrary functions by polynomials, Simpson's rule for numerical integration, and Euler's method for the solution of initial value problems. However, the main breakthrough of numerical methods is connected with the advances in computer technology made within the last four decades. Only the electronic computer allows one to perform extensive numerical computations without error and within a reasonable amount of time. Hence, progress in numerical analysis and computer science have always been closely interrelated in recent history.

This book will introduce the reader to the following branches of numerical analysis:

Solution of systems of linear and nonlinear equations,
Numerical solution of matrix eigenvalue problems,
Interpolation and numerical integration,
Numerical solution of initial and boundary value problems for differential equations,
Numerical solution of integral equations.

Of course, in an introductory exposition of only about three hundred pages it is impossible to cover all of these areas exhaustively. Therefore, the reader should not expect a comprehensive treatment of all existing numerical procedures. As already pointed out in the preface, our goal will be to guide the reader toward the basic ideas and questions in each of the above topics with an emphasis on the analysis and the understanding of numerical methods rather than merely their description. In order to achieve this, we will try to illustrate general principles by way of considering the main and most important methods, and we will leave aside discussions of more elaborate details of advanced methods and the consideration of lengthy subtleties for exceptional cases. Given the rapid development of numerical methods, a reasonable introduction to numerical analysis has to confine itself to presenting a solid foundation by restricting the presentation to the basic principles and procedures.

The book includes a chapter on the necessary basic functional-analytic tools for the solid mathematical foundation of numerical analysis. These are indispensable for any deeper study and understanding of numerical methods, in particular for differential equations and integral equations.

The limit of space and the taste and restrictions in experience of the author have caused the omission of some important topics such as linear and nonlinear optimization, approximation theory, and parallel computing, among others. On the other hand, with separate chapters on the solution of ill-conditioned systems of linear equations and the numerical solution of integral equations two topics are included that do not appear in most introductions to numerical analysis. They are included because of their importance and in order to indicate to the reader where the author's mathematical research interests lie.

A study of numerical analysis remains incomplete without the numerical experience of individually implementing the numerical algorithms. It

is very important to build up a familiarity with numerical methods by actually seeing the numbers working. For example, one has to complement the theoretical understanding of the method of successive approximations by the experience of actually running the numerical schemes. After having understood the basic principles of a numerical method, it is important to develop the ability to actually implement the method numerically and work with it. In this sense the reader is encouraged to test on the computer numerically all of the algorithms presented in this book.

The organization of the book is as follows. The first part of the book, Chapters 2 to 7, covers numerical linear algebra and is concerned with the solution of systems of linear and nonlinear equations. The necessary functional-analytic tools will be presented in Chapter 3. The second part of the book, Chapters 8 to 12, covers numerical analysis and is concerned with interpolation, numerical integration, and the numerical solution of differential and integral equations. At the reader's convenience it is possible to study most of the second part of the book before reading the first part, with the exception of the chapter on functional analysis. Each chapter concludes with a set of problems. These are intended as exercises and applications of the material given in the chapter.

The references at the end of the book are intended as a possible guide to some of the literature covering the topics of the individual chapters more exhaustively. The list of references is not meant as a bibliography on the vast number of introductions to numerical analysis competing with this book. However, we explicitly encourage the reader to explore the libraries and consult some of the other volumes on numerical analysis in order to develop a broad perspective.

2
Linear Systems

The solution of systems of linear equations arises in various parts of mathematics and is of central importance in numerical analysis. To illustrate the significance of linear systems, we will start this chapter by providing some examples of their occurrence as part of the numerical solution of differential and integral equations. After seeing the examples, we will proceed with the solution of systems of linear equations. In principle, we have to distinguish between two groups of methods for the solution of linear systems:

1. In the so-called *direct methods*, or *elimination methods*, the exact solution, in principle, is determined through a finite number of arithmetic operations (in real arithmetic leaving aside the influence of roundoff errors).
2. In contrast to this, *iterative methods* generate a sequence of approximations to the solution by repeating the application of the same computational procedure at each step of the iteration. Usually, they are applied for large systems with special structures that ensure convergence of the successive approximations.

A key consideration for the selection of a solution method for a linear system is its structure. In some problems, the matrix of the linear system may be a full matrix, i.e., it has few zero entries. And in other problems, the matrix may be very large and sparse, i.e., only a small fraction of the entries are different from zero. Roughly speaking, direct methods are best for full matrices, whereas iterative methods are best for very large and sparse matrices.

We will begin our treatment of linear systems by presenting the best-known and most widely used direct method, which is attributed to Gauss, since it is based on considerations published by Gauss in 1801 in his *Disquisitiones Arithmeticae.* The chapter concludes with a brief description of elimination by orthonormal decomposition.

In this book, for an $m \times n$ matrix $A = (a_{jk})$, $j = 1, \dots, m$, $k = 1, \dots, n$, with real or complex coefficients, A^T shall always denote the *transposed* matrix; i.e., A^T is the $n \times m$ matrix with entries

$$a^T_{kj} = a_{jk}, \quad k = 1, \dots, n, \; j = 1, \dots, m.$$

By A^* we denote the *adjoint* of the matrix A; i.e., $A^* = \overline{A}^T$ is the transpose of the matrix with complex conjugate entries. In particular, the transpose and adjoint of a row vector are column vectors and vice versa.

2.1 Examples for Systems of Equations

Example 2.1 We consider the discretization of the boundary value problem for the ordinary differential equation

$$-u''(x) = f(x, u(x)), \quad x \in [0,1], \tag{2.1}$$

with boundary condition

$$u(0) = u(1) = 0. \tag{2.2}$$

Here, $f : [0,1] \times \mathbb{R} \to \mathbb{R}$ is a given continuous function, and we are looking for a twice continuously differentiable solution $u : [0,1] \to \mathbb{R}$. Boundary value problems of this type occur, for example, in the mathematical treatment of vibrations of a string or a rod and in the solution of heat conduction problems. They often also arise in the solution of problems like the following Example 2.2 after applying separation of variables. The theory of ordinary differential equations (see [12]) provides conditions on the right-hand side f of (2.1), ensuring existence and uniqueness of a solution u to the boundary value problem (2.1)–(2.2) (for the case of linear differential equations see also Chapter 11).

For the approximate solution we choose an equidistant subdivision of the interval $[0,1]$ by setting

$$x_j = jh, \quad j = 0, \dots, n+1,$$

where the step size is given by $h = 1/(n+1)$ with $n \in \mathbb{N}$. At the internal grid points x_j, $j = 1, \dots, n$, we replace the differential quotient in the differential equation (2.1) by the difference quotient

$$u''(x_j) \approx \frac{1}{h^2}\left[u(x_{j+1}) - 2u(x_j) + u(x_{j-1})\right]$$

to obtain the system of equations

$$-\frac{1}{h^2}\left[u_{j-1} - 2u_j + u_{j+1}\right] = f(x_j, u_j), \quad j = 1, \ldots, n,$$

for approximate values u_j to the exact solution $u(x_j)$. This system has to be complemented by the two boundary conditions $u_0 = u_{n+1} = 0$. For an abbreviated notation we introduce the $n \times n$ matrix

$$A = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}$$

and the vectors $U = (u_1, \ldots, u_n)^T$ and $F(U) = (f(x_1, u_1), \ldots, f(x_n, u_n))^T$. Then our system of equations, including the boundary conditions, reads

$$AU = F(U). \tag{2.3}$$

For obvious reasons, the above matrix A is called a *tridiagonal* matrix, and the vector F is diagonal; i.e., the jth component of F depends only on the jth component of u. If (2.1) is a linear differential equation, i.e., if f depends linearly on the second variable u, then the tridiagonal system of equations (2.3) also is linear.

The following two questions will be addressed later in the book (see Chapter 11):

1. Can we establish existence and uniqueness of a solution to the system of equations (2.3) for sufficiently small step size h, provided that the boundary value problem (2.1)–(2.2) itself is uniquely solvable?
2. How large is the error between the approximate solution u_j and the exact solution $u(x_j)$? Do we have convergence of the approximate solution towards the exact solution as $h \to 0$?

At this point we would like only to point out that the discretization of boundary value problems for ordinary differential equations leads to systems of equations with a large number of unknowns, since we expect that in order to achieve a reasonably accurate approximation we need to choose the step size h sufficiently small. □

Example 2.2 We now consider the discretization of the boundary value problem for the elliptic partial differential equation

$$-\Delta\, u(x) = f(x, u(x)), \quad x \in D, \tag{2.4}$$

with Dirichlet boundary condition

$$u(x) = 0, \quad x \in \partial D. \tag{2.5}$$

Here, $D \subset \mathbb{R}^2$ is a bounded domain, Δ denotes the Laplacian

$$\Delta u := \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2},$$

$f : D \times \mathbb{R} \to \mathbb{R}$ is a given continuous function, and we are looking for a solution $u : \bar{D} \to \mathbb{R}$ that is continuous in $\bar{D}$ and twice continuously differentiable in D. Boundary value problems of this type arise, for example, in potential theory and in heat conduction problems. The theory of elliptic partial differential equations (see [24]) provides conditions on the given function f that ensure existence and uniqueness of a solution u.

For describing a numerical approximation method we restrict ourselves to the case of the square $D = (0,1) \times (0,1)$. We choose an equidistant quadratic grid with grid points

$$x_{ij} = (ih, jh), \quad i, j = 0, \dots, n+1,$$

where the step size again is given by $h = 1/(n+1)$ with $n \in \mathbb{N}$. Analogously to the previous example, at the internal grid points x_{ij}, $i, j = 1, \dots, n$, we replace the Laplacian by the Laplace difference operator

$$\Delta u(x_{i,j}) \approx \frac{1}{h^2}\left[u(x_{i+1,j}) + u(x_{i-1,j}) + u(x_{i,j+1}) + u(x_{i,j-1}) - 4u(x_{ij})\right].$$

Obviously, for each point x_{ij}, this difference operator has nonvanishing weights only at the four neighboring points on the vertical and horizontal line through x_{ij}. This observation also illustrates why the set of grid points with nonvanishing weights is called the star associated with the Laplace difference operator. Using this difference approximation leads to the system of equations

$$\frac{1}{h^2}\left[4u_{ij} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1}\right] = f(x_{ij}, u_{ij}), \quad i, j = 1, \dots, n,$$

for approximate values u_{ij} to the exact solution $u(x_{ij})$. This system has to be complemented by the boundary conditions

$$u_{0,j} = u_{n+1,j} = 0, \quad j = 0, \dots, n+1,$$

at the grid points on the vertical parts and

$$u_{i,0} = u_{i,n+1} = 0, \quad i = 1, \dots, n,$$

at the grid points on the horizontal parts of the boundary ∂D. In order to write this system in matrix form we rearrange the unknowns by ordering them row by row and setting

$$u_1 = u_{11},\ u_2 = u_{21}, \dots, u_n = u_{n,1},\ u_{n+1} = u_{12}, \dots, u_m = u_{nn},$$

where $m = n^2$. Furthermore, we introduce an $m \times m$ matrix A in the form of an $n \times n$ block tridiagonal matrix

$$A = \frac{1}{h^2} \begin{pmatrix} B & -I & & & & \\ -I & B & -I & & & \\ & -I & B & -I & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & -I & B & -I \\ & & & & -I & B \end{pmatrix},$$

where I denotes the $n \times n$ identity matrix and B is the $n \times n$ tridiagonal matrix

$$B = \begin{pmatrix} 4 & -1 & & & & \\ -1 & 4 & -1 & & & \\ & -1 & 4 & -1 & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & -1 & 4 & -1 \\ & & & & -1 & 4 \end{pmatrix}.$$

After introducing the vectors U and $F(U)$ analogously to Example 2.1, we can rewrite the system of equations in the short form

$$AU = F(U), \tag{2.6}$$

which also includes the boundary conditions.

Again we postpone the questions of unique solvability of the system (2.6) and the problem of convergence and error estimates for later parts of the book (see Chapter 11). Here, we conclude the example with the observation that the system has n^2 unknowns, where n will be fairly large if the step size h is sufficiently small in order to achieve a reasonably accurate approximation to the solution of the boundary value problem. These large systems of equations arising in the discretization of partial differential equations call for efficient solution methods. □

Example 2.3 Consider the linear integral equation

$$\varphi(x) - \int_0^1 K(x,y)\varphi(y)\,dy = f(x), \quad x \in [0,1],$$

where $K : [0,1] \times [0,1] \to \mathbb{R}$ and $f : [0,1] \to \mathbb{R}$ are given continuous functions and where we seek a continuous solution $\varphi : [0,1] \to \mathbb{R}$. Such integral equations either arise directly in the solution of applied problems, or more often they occur indirectly in the solution of boundary value problems for differential equations. If the homogeneous form of this equation, i.e., the integral equation with the right-hand side $f = 0$, admits only the trivial solution $\varphi = 0$, then for each f the inhomogeneous integral equation has a unique solution φ (see Chapter 12).

For the numerical approximation we replace the integral by the rectangular sum

$$\int_0^1 K(x,y)\varphi(y)\,dy \approx \frac{1}{n}\sum_{k=1}^{n} K(x,x_k)\varphi(x_k)$$

with equidistant grid points $x_k = k/n$, $k = 1,\dots,n$. If we require the approximated equation to be satisfied only at the grid points, we arrive at the system of linear equations

$$\varphi_j - \frac{1}{n}\sum_{k=1}^{n} K(x_j,x_k)\varphi_k = f(x_j), \quad j = 1,\dots,n,$$

for approximate values φ_j to the exact solution $\varphi(x_j)$. As in the preceding examples, we postpone the question of unique solvability of the linear system and the convergence and error analysis (see Chapter 12). □

Example 2.4 In this last example we will briefly touch on the *method of least squares.* Consider some (physical) quantity u depending on time t and a parameter vector $a = (a_1,\dots,a_n)^T \in \mathbb{R}^n$ in terms of a known function

$$u(t) = f(t;a).$$

In order to determine the values of the parameter a (representing some physical constants), one can take m measurements of u at different times $t_1,\dots,t_m$ and then try to find a by solving the system of equations

$$u(t_j) = f(t_j;a), \quad j = 1,\dots,m.$$

If $m = n$, this system consists of n equations for the n unknowns $a_1,\dots,a_n$. However, in general, the measurements will be contaminated by errors. Therefore, usually one will take $m > n$ measurements and then will try to determine a by requiring the deviations

$$u(t_j) - f(t_j;a), \quad j = 1,\dots,m,$$

to be as small as possible. Usually the latter requirement is posed in the least squares sense, i.e., the parameter a is chosen such that

$$g(a) := \sum_{k=1}^{m} [u(t_k) - f(t_k;a)]^2$$

attains a minimal value. The necessary conditions for a minimum,

$$\frac{\partial g}{\partial a_j} = 0, \quad j = 1,\dots,n,$$

lead to the *normal equations*

$$\sum_{k=1}^{m} [u(t_k) - f(t_k;a)]\,\frac{\partial f(t_k;a)}{\partial a_j} = 0, \quad j = 1,\dots,n,$$

for the method of least squares. These constitute a system of n, in general, nonlinear equations for the n unknowns $a_1, \dots, a_n$. □

At this point, the reader should be convinced of the need for effective methods for solving large systems of linear and nonlinear equations and be willing to be introduced to such methods in the subsequent chapters. We also wish to note that the discretization of differential equations leads to sparse matrices, whereas for the least squares problem and the discretization of integral equations one is faced with full matrices.

2.2 Gaussian Elimination

We proceed with describing the *Gaussian elimination method* for a *system of linear equations*

$$Ax = y.$$

Here A is a given $n \times n$ matrix $A = (a_{jk})$ with real (or complex) entries, y a given right-hand side $y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ (or $\mathbb{C}^n$), and we are looking for a solution vector $x = (x_1, \dots, x_n)^T \in \mathbb{R}^n$ (or $\mathbb{C}^n$). More explicitly, our system of equations can be written in the form

$$\sum_{k=1}^{n} a_{jk} x_k = y_j, \quad j = 1, \dots, n;$$

that is,

$$\begin{array}{ccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n = y_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n = y_2 \\ & & & & \cdots & & \\ a_{n1}x_1 & + & a_{n2}x_2 & + & \cdots & + & a_{nn}x_n = y_n. \end{array}$$

Assuming that the reader is familiar with basic linear algebra, we recall the following various ways of saying that the matrix A is nonsingular:

1. The inverse matrix A^{-1} exists.
2. For each y the linear system $Ax = y$ has a unique solution.
3. The homogeneous system $Ax = 0$ has only the trivial solution.
4. The determinant of A satisfies $\det A \neq 0$.
5. The rows (columns) of A are linearly independent.

The very basic idea of the Gaussian elimination method is to use the first equation to eliminate the first unknown from the last $n-1$ equations, then use the new second equation to eliminate the second unknown from the last $n-2$ equations, etc. This way, by $n-1$ such eliminations the given linear

system is transformed into an equivalent linear system that is of *triangular* form

$$\begin{aligned}
b_{11}x_1 + b_{12}x_2 + \quad \cdots \quad + \quad b_{1n}x_n &= z_1\\
b_{22}x_2 + \quad \cdots \quad + \quad b_{2n}x_n &= z_2\\
\cdots\qquad&\\
b_{n-1,n-1}x_{n-1} + b_{n-1,n}x_n &= z_{n-1}\\
b_{nn}x_n &= z_n
\end{aligned}$$

Recall that two linear systems are called *equivalent* if every solution of one is a solution of the other. The triangular system can be solved recursively by first obtaining x_n from the last equation, then obtaining x_{n-1} from the second to last equation, etc. This procedure is known as *backward substitution*. Explicitly, it is described by $x_n = z_n/b_{nn}$ and

$$x_m = \frac{1}{b_{mm}}\left(z_m - \sum_{k=m+1}^{n} b_{m,k}x_k\right), \quad m = n-1, n-2, \dots, 1.$$

We begin by considering a nonsingular matrix A. To eliminate the unknown x_1, for $j = 2, \dots, n$ we multiply the first equation by a_{j1}/a_{11} and subtract the result from the jth equation. For this we have to require that $a_{11} \neq 0$. Since we assume the matrix to be nonsingular, this can be achieved by reordering the rows or the columns of the given system. This procedure leads to a system of the form

$$\begin{aligned}
b_{11}x_1 + b_{12}x_2 + \quad \cdots \quad + b_{1n}x_n &= z_1\\
a_{22}^{(2)}x_2 + \quad \cdots \quad + a_{2n}^{(2)}x_n &= y_2^{(2)}\\
\cdots\qquad&\\
a_{n2}^{(2)}x_2 + \quad \cdots \quad + a_{nn}^{(2)}x_n &= y_n^{(2)}
\end{aligned}$$

with the new coefficients given by

$$b_{1k} := a_{1k}^{(1)}, \qquad k = 1, \dots, n,$$

$$a_{jk}^{(2)} := a_{jk}^{(1)} - \frac{a_{j1}^{(1)}a_{1k}^{(1)}}{a_{11}^{(1)}}, \qquad j, k = 2, \dots, n,$$

and the new right-hand sides given by

$$z_1 := y_1^{(1)}, \quad y_j^{(2)} := y_j^{(1)} - \frac{a_{j1}^{(1)}y_1^{(1)}}{a_{11}^{(1)}}, \quad j = 2, \dots, n.$$

Here, for the coefficients and right-hand sides of the original system we have set $a_{jk}^{(1)} := a_{jk}$ and $y_j^{(1)} := y_j$.

Proceeding in this way, the given $n \times n$ system for the unknowns $x_1, \dots, x_n$ is equivalently transformed into an $(n-1) \times (n-1)$ system for the unknowns $x_2, \dots, x_n$. Adding a multiple of one row of a matrix to another row does not change the value of its determinant. Therefore, in the above elimination the determinant of the system remains the same (with the exception of a possible change of its sign if the order of rows or columns is changed). Hence, the resulting $(n-1) \times (n-1)$ system for $x_2, \dots, x_n$ again has a nonvanishing determinant, and we can apply precisely the same procedure to eliminate the second unknown x_2 from the remaining $(n-1) \times (n-1)$ system.

By repeating this process we complete the *forward elimination*, by which the system of linear equations

$$\begin{aligned}
a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + \cdots + a_{1n}^{(1)} x_n &= y_1^{(1)} \\
a_{21}^{(1)} x_1 + a_{22}^{(1)} x_2 + \cdots + a_{2n}^{(1)} x_n &= y_2^{(1)} \\
&\cdots \\
a_{n1}^{(1)} x_1 + a_{n2}^{(1)} x_2 + \cdots + a_{nn}^{(1)} x_n &= y_n^{(1)}
\end{aligned}$$

with a nonsingular matrix $A = (a_{jk}^{(1)})$ is equivalently transformed into a triangular system

$$\begin{aligned}
b_{11} x_1 + b_{12} x_2 + \cdots + b_{1n} x_n &= z_1 \\
b_{22} x_2 + \cdots + b_{2n} x_n &= z_2 \\
&\cdots \\
b_{n-1,n-1} x_{n-1} + b_{n-1,n} x_n &= z_{n-1} \\
b_{nn} x_n &= z_n
\end{aligned}$$

by $n-1$ recursive elimination steps of the form

$$\left.\begin{aligned}
a_{jk}^{(m+1)} &:= a_{jk}^{(m)} - \frac{a_{jm}^{(m)} a_{mk}^{(m)}}{a_{mm}^{(m)}}, \quad j, k = m+1, \dots, n, \\
y_j^{(m+1)} &:= y_j^{(m)} - \frac{a_{jm}^{(m)} y_m^{(m)}}{a_{mm}^{(m)}}, \quad j = m+1, \dots, n,
\end{aligned}\right\} \quad m = 1, \dots, n-1.$$

The coefficients and the right-hand sides of the final triangular system are given by

$$b_{jk} := a_{jk}^{(j)}, \quad k = j, \ldots, n, \; j = 1, \ldots, n,$$

and

$$z_j := y_j^{(j)}, \quad j = 1, \ldots, n.$$

The condition $a_{mm}^{(m)} \neq 0$, which is necessary for performing the algorithm, always can be achieved by a reordering of the rows or columns, since otherwise the matrix A would not be nonsingular.

We would like to compress the operations of one elimination step into the following scheme

a	b
c	d

where the rectangle illustrates the remaining part of the matrix and the right-hand side for which the elimination has to be performed. Here, a stands for the elimination element, or *pivot element;* the elements b in the elimination row remain unchanged; the elements c of the elimination column are replaced by zero (with the exception of the pivot element a); and the remaining elements d are changed according to the rule

$$d \rightarrow d - \frac{bc}{a}.$$

We note that in computer calculations, of course, the new values for the coefficients of the matrix and the right-hand sides can be stored in the locations held by the old values.

More explicitly, the entire Gaussian elimination can be written in the following algorithmic form.

Algorithm 2.5 *(Gaussian elimination)*

1. Forward elimination:

For $m = 1, \ldots, n-1$ *do*

  for $j = m+1, \ldots, n$ *do*

    for $k = m+1, \ldots, n$ *do* $a_{jk} := a_{jk} - \dfrac{a_{jm} a_{mk}}{a_{mm}}$

    $y_j := y_j - \dfrac{a_{jm} y_m}{a_{mm}}$

2. Backward substitution:

For $m = n, n-1, \ldots, 1$ *do* $x_m := y_m$

for $k = m+1, \ldots, n$ *do* $x_m := x_m - a_{mk} x_k$

$$x_m := \frac{x_m}{a_{mm}}$$

If the matrix A is singular and has rank r, the elimination procedure will terminate after r steps. The matrix of the remaining $(n-r) \times (n-r)$ system for the unknowns $x_{r+1}, \ldots, x_n$ is the zero matrix, because otherwise the rank of A would be different from r. Hence, in this case the given linear system is solvable if and only if the right-hand sides after r elimination steps satisfy

$$z_{r+1} = \cdots = z_n = 0.$$

The solutions can be found from the triangular system by arbitrarily choosing $x_{r+1}, \ldots, x_n$ and then recursively determining $x_r, \ldots, x_0$. This way we obtain the $(n-r)$-dimensional solution manifold.

In order to control the influence of roundoff errors we want to keep the quotient $a_{jm}^{(m)}/a_{mm}^{(m)}$ small; i.e., we want to have a large pivot element $a_{mm}^{(m)}$. Therefore, instead of only requiring $a_{mm}^{(m)} \neq 0$, in practice, either *complete pivoting* or *partial row or column pivoting* is employed. For complete pivoting, both the rows and the columns are reordered such that $a_{mm}^{(m)}$ has maximal absolute value in the $(n-m+1) \times (n-m+1)$ matrix remaining for the mth forward elimination step. In order to minimize the additional computational cost caused by pivoting, for row (or column) pivoting the rows (or columns) are reordered such that $a_{mm}^{(m)}$ has maximal absolute value in the elimination column (or row), i.e., in the mth column (or row). Of course, in the actual implementation of the Gaussian elimination algorithm the reordering of rows and columns need not be done explicitly. Instead, the interchange may be done only implicitly by leaving the pivot element at its original location and keeping track of the interchange of rows and columns through the associated permutation matrix.

The following example illustrates that partial pivoting does not always prevent loss of accuracy in the numerical computations.

Example 2.6 We consider the system

$$x_1 + 200x_2 = 100$$

$$x_1 + \quad x_2 = 1$$

with the exact solution $x_1 = 100/199 = 0.502\ldots$, $x_2 = 99/199 = 0.497\ldots$. For the following computations we use two-decimal-digit floating-point

arithmetic. Column pivoting leads to a_{11} as pivot element, and the elimination yields

$$\begin{aligned} x_1 + 200x_2 &= 100 \\ -\ 200x_2 &= -99, \end{aligned}$$

since $199 = 200$ in two-digit floating-point representation. From the second equation we then have $x_2 = 0.50$ ($0.495 = 0.50$ in two decimal digits), and from the first equation it finally follows that $x_1 = 0$.

However, if by complete pivoting we choose a_{12} as pivot element, the elimination leads to

$$\begin{aligned} x_1 + 200x_2 &= 100 \\ x_1 \qquad\quad &= 0.5 \end{aligned}$$

($0.995 = 1.00$ in two decimal digits), and from this we get the solution $x_1 = 0.5$, $x_2 = 0.5$ ($0.4975 = 0.50$ in two decimal digits), which is correct to two decimal digits. □

Since complete pivoting is more costly than partial pivoting, in practical computations one can try to overcome the disadvantages of partial pivoting by *scaling* the matrix. This means that if $B = D_1AD_2$, in order to obtain the solution x of $Ax = y$ we first solve $Bz = D_1y$ for z and then determine x from $x = D_2z$. Here D_1 and D_2 are some diagonal matrices chosen such that for the matrix B the row and column sums of the absolute values are approximately equal. A *diagonal matrix* $D = (d_{jk})$ is a matrix with the off-diagonal elements equal to zero; i.e., $d_{jk} = 0$ for $j \neq k$. For a detailed discussion of scaling we refer to [27]. Unfortunately, there is no known general procedure for such scaling, i.e., for choosing the diagonal matrices D_1 and D_2.

For an estimate of the computational cost of Gaussian elimination we perform a count of the number of multiplications. By α_n we denote the number of multiplications that are required for solving a triangular $n \times n$ system by back substitution. Obviously, for α_n we have the recurrence relation

$$\alpha_n = \alpha_{n-1} + n,$$

since we need n multiplications to obtain x_1 from the first equation after having already determined $x_2, \ldots, x_n$. Hence, we have

$$\alpha_n = \sum_{k=1}^{n} k = \frac{n(n+1)}{2},$$

since $\alpha_1 = 1$. By $\beta_{n,r}$ we denote the number of multiplications needed for the forward elimination simultaneously for r different right-hand sides. Here we have the recurrence relation

$$\beta_{n,r} = \beta_{n-1,r} + (n + r)(n - 1),$$

since the elimination of the unknown x_1 requires $n + r$ multiplications for each row of the $n - 1$ rows. From this it follows that

$$\beta_{n,r} = \sum_{k=1}^{n}(k + r)(k - 1) = \frac{n^3}{3} - \frac{n}{3} + \frac{n(n-1)r}{2}$$

because $\beta_{1,r} = 0$. Adding $r\alpha_n$ and $\beta_{n,r}$ we obtain the following result.

Theorem 2.7 *Gaussian elimination for the simultaneous solution of an $n \times n$ system for r different right-hand sides requires a total of*

$$\frac{n^3}{3} + rn^2 - \frac{n}{3}$$

multiplications.

The computational cost, counting only the multiplications, in Gaussian elimination is $n^3/3 + O(n^2)$. It is left to the reader to show that the number of additions is also $n^3/3 + O(n^2)$ (see Problem 2.7). Doubling the number of unknowns increases the computation time by a factor of eight. Assuming $1\,\mu\,\text{sec} = 10^{-6}$ sec per addition and multiplication, i.e., on a computer with one million floating point operations per second, the solution of a system with $n = 10^3$ requires approximately ten minutes, and with $n = 10^4$ it requires approximately six days. This illustrates dramatically that for the solution of large linear systems iterative methods, which we will study in Chapter 4, are better suited than direct methods. Row or column pivoting leads to an additional cost proportional to n^2, whereas complete pivoting adds costs proportional to n^3. For the latter reason, complete pivoting is used only rarely in practical computations.

The Gaussian algorithm also allows the computation of the determinant and the inverse of a matrix A. The determinant $\det A$ is simply given by the product of the diagonal elements in the triangular matrix obtained through the elimination procedure. If the determinant is computed using expansions by submatrices, then the operational count is $n!$ multiplications, as compared to $n^3/3$ for Gaussian elimination. This illustrates why Cramer's rule for the solution of linear systems is only a theoretical mathematical tool and not a tool for practical computations.

The inverse of a matrix is obtained by solving the linear system simultaneously for the n right-hand sides given by the columns of the identity matrix, i.e., by solving the n systems

$$Ax_i = e_i, \quad i = 1, \ldots, n,$$

where e_i is the ith column of the identity matrix. Then the n solutions $x_1, \ldots, x_n$ will provide the columns of the inverse matrix A^{-1}. We would like to stress that one does not want to solve a system $Ax = y$ by first

computing A^{-1} and then evaluating $x = A^{-1}y$, since this generally leads to considerably higher computational costs.

The *Gauss–Jordan method* is an elimination algorithm that in each step eliminates the unknown both above and below the diagonal. The complete elimination procedure transforms the system equivalently into a diagonal system. The multiplication count shows a computational cost of order $n^3/2 + O(n^2)$, i.e., an increase of 50 percent over Gaussian elimination. Hence, the Gauss–Jordan method is rarely used in applications. For details we refer to [26, 27].

2.3 LR Decomposition

In the sequel we will indicate how Gaussian elimination provides an *LR decomposition* (or *factorization*) of a given matrix.

Definition 2.8 *A factorization of a matrix A into a product*

$$A = LR$$

of a lower (left) triangular matrix L and an upper (right) triangular matrix R is called an LR decomposition *of A.*

A matrix $A = (a_{jk})$ is called *lower triangular* or *left triangular* if $a_{jk} = 0$ for $j < k$; it is called *upper triangular* or *right triangular* if $a_{jk} = 0$ for $j > k$. The product of two lower (upper) triangular matrices again is lower (upper) triangular, lower (upper) triangular matrices with nonvanishing diagonal elements are nonsingular, and the inverse matrix of a lower (upper) triangular matrix again is lower (upper) triangular (see Problem 2.14).

Theorem 2.9 *For a nonsingular matrix A, Gaussian elimination (without reordering rows and columns) yields an LR decomposition.*

Proof. In the first elimination step we multiply the first equation by a_{j1}/a_{11} and subtract the result from the jth equation; i.e., the matrix $A_1 = A$ is multiplied from the left by the lower triangular matrix

$$L_1 = \begin{pmatrix} 1 & & & & \\ -\dfrac{a_{21}}{a_{11}} & 1 & & & \\ \cdot & & \cdot & \cdot & \cdot \\ -\dfrac{a_{n1}}{a_{11}} & & & & 1 \end{pmatrix}.$$

The resulting matrix $A_2 = L_1 A_1$ is of the form

$$A_2 = \begin{pmatrix} a_{11} & * \\ 0 & \tilde{A}_{n-1} \end{pmatrix},$$

where $\tilde{A}_{n-1}$ is an $(n-1) \times (n-1)$ matrix. In the second step the same procedure is repeated for the $(n-1) \times (n-1)$ matrix $\tilde{A}_{n-1}$. The corresponding $(n-1) \times (n-1)$ elimination matrix is completed as an $n \times n$ triangular matrix L_2 by setting the diagonal element in the first row equal to one. In this way, $n-1$ elimination steps lead to

$$L_{n-1} \cdots L_1 A = R,$$

with nonsingular lower triangular matrices $L_1, \ldots, L_{n-1}$ and an upper triangular matrix R. From this we find

$$A = LR,$$

where L denotes the inverse of the product $L_{n-1} \cdots L_1$. □

We wish to point out that not every nonsingular matrix allows an LR decomposition. For example,

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

has no LR decomposition. However, since Gaussian elimination with row reordering always works, for each nonsingular matrix A there exists a permutation matrix P such that PA has an LR decomposition (see Problem 2.16). A *permutation matrix* is a matrix of the form $P = (e_{p(1)}, \ldots, e_{p(n)})$ where $e_1, \ldots, e_n$ are the columns of the identity matrix and $p(1), \ldots, p(n)$ is a permuation of $1, \ldots, n$.

Recall that an $n \times n$ matrix A is called *symmetric* if it has real coefficients and $A = A^T$. A symmetric matrix A is called *positive definite* if $x^T Ax > 0$ for all $x \in \mathbb{R}^n$ with $x \neq 0$. Positive definite matrices have positive diagonal elements (see Problem 2.10), and therefore a reordering of rows and columns is not necessary for Gaussian elimination (for pivoting, the largest diagonal element is chosen). It can be shown (see Problem 2.13) that symmetry and positive definiteness are preserved throughout the elimination if diagonal elements are taken as pivot elements. Therefore, for symmetric positive definite matrices the LR decomposition is always possible. If $A = LR$, then we have also $A = A^T = R^T L^T$, and from Problem 2.15 we can deduce that L can be normalized such that $A = LL^T$. Such a decomposition is used in the *Cholesky method* for the solution of linear systems with symmetric positive definite matrices. Because of symmetry, the computational cost for the Cholesky method is $n^3/6 + O(n^2)$ multiplications and $n^3/6 + O(n^2)$ additions. For details we refer to [26, 27].

2.4 QR Decomposition

We conclude this chapter by describing a second elimination method for linear systems, which leads to a *QR decomposition.*

Definition 2.10 *A factorization of a matrix A into a product*

$$A = QR$$

of a unitary matrix Q and an upper (right) triangular matrix R is called a QR decomposition *of A.*

We recall that a matrix Q is called *unitary* if

$$QQ^* = Q^*Q = I.$$

The product of two unitary matrices again is unitary.

In terms of the columns of the matrices $A = (a_1, \dots, a_n)$ and $Q = (q_1, \dots, q_n)$ and the coefficients of $R = (r_{jk})$, the QR decomposition $A = QR$ means that

$$a_k = \sum_{i=1}^{k} r_{ik} q_i, \quad k = 1, \dots, n. \tag{2.7}$$

Hence, the vectors $a_1, \dots, a_n$ of $\mathbb{C}^n$ have to be orthonormalized from the left to the right into an orthonormal basis $q_1, \dots, q_n$. This, for example, can be achieved by the Gram–Schmidt orthonormalization procedure (see Theorem 3.18). However, since the Gram–Schmidt orthonormalization tends to be numerically unstable, we describe the QR decomposition by Householder matrices.

Definition 2.11 *A matrix H of the form*

$$H = I - 2vv^*,$$

where v is column vector with $v^*v = 1$, i.e., a unit vector, is called a Householder matrix.

Remark 2.12 *Householder matrices are unitary and satisfy $H = H^*$.*

Proof. We compute

$$H^* = I^* - 2(vv^*)^* = I - 2vv^* = H$$

and

$$HH^* = H^*H = (I - 2vv^*)(I - 2vv^*) = I - 4vv^* + 4vv^*vv^* = I,$$

where we use that $v^*v = 1$. □

Geometrically a Householder matrix corresponds to reflection across the plane through the origin orthogonal to v. To see this we write

$$x = vv^*x + y$$

with the component vv^*x of $x \in \mathbb{C}^n$ in the v-direction and a component y orthogonal to v. Then we obtain

$$Hx = x - 2vv^*x = -vv^*x + y;$$

i.e., Hx has the opposite component $-vv^*x$ in the v-direction and the same component y orthogonal to v. Because of this property, Householder matrices are also called elementary reflection matrices.

We now describe the elimination of the unknown x_1 by multiplying A from the left by a Householder matrix $H_1 = I - 2v_1v_1^*$. By a_1 we denote the first column of A and by e_k the kth column of the identity matrix; in particular, $e_1 = (1, 0, \dots, 0)^*$. Then the first column b_1 of the product H_1A is given by

$$b_1 = H_1Ae_1 = H_1a_1 = a_1 - 2v_1v_1^*a_1.$$

We would like to achieve that $b_1 = \sigma e_1$ with $\sigma \neq 0$. Hence, except for the first row, v_1 must be a multiple of a_1. Therefore, we try

$$u_1 = a_1 \mp \sigma e_1 \tag{2.8}$$

with

$$\sigma = \begin{cases} \dfrac{a_{11}}{|a_{11}|}\sqrt{a_1^*a_1}\,, & a_{11} \neq 0, \\[2ex] \sqrt{a_1^*a_1}\,, & a_{11} = 0. \end{cases}$$

Then we have

$$u_1^*u_1 = 2(a_1^*a_1 \mp |a_{11}|\sqrt{a_1^*a_1}\,)$$

and

$$u_1^*a_1 = a_1^*a_1 \mp |a_{11}|\sqrt{a_1^*a_1} = \frac{1}{2}\,u_1^*u_1.$$

Without loss of generality we may assume that $\sqrt{a_1^*a_1} - |a_{11}| > 0$, since otherwise we would have that $a_1 = a_{11}e_1$, i.e., that the first column already has the required form. Therefore, if we finally choose

$$v_1 = \frac{u_1}{\sqrt{u_1^*u_1}}\,,$$

then v_1 is a unit vector, and as requested we have

$$b_1 = a_1 - \frac{2}{u_1^*u_1}\,u_1u_1^*a_1 = a_1 - u_1 = \pm\sigma e_1.$$

The remaining columns $b_k = H_1Ae_k$ are obtained from the columns a_k of A by

$$b_k = H_1Ae_k = H_1a_k = a_k - 2v_1v_1^*a_k = a_k - \frac{u_1^*a_k}{u_1^*a_1}\,u_1, \quad k = 2, \dots, n.$$

From the two possible signs in (2.8) the positive sign yields the numerically more stable variant.

The same procedure is now repeated for the remaining $(n-1) \times (n-1)$ matrix. The corresponding $(n-1) \times (n-1)$ Householder matrix has to be completed as an $n \times n$ Householder matrix. In general, if A_k is an $n \times n$ matrix of the form

$$A_k = \begin{pmatrix} R_k & * \\ 0 & \tilde{A}_{n-k} \end{pmatrix}$$

with a $k \times k$ upper triangular matrix R_k and an $(n-k) \times (n-k)$ matrix $\tilde{A}_{n-k}$, we apply the Householder transformation described above with the first column of $\tilde{A}_{n-k}$. With the corresponding $(n-k) \times (n-k)$ Householder matrix $\tilde{H}_{n-k}$ the $n \times n$ matrix

$$H_k = \begin{pmatrix} I_k & 0 \\ 0 & \tilde{H}_{n-k} \end{pmatrix}$$

yields an $n \times n$–Householder matrix H_k that leaves the first k columns in triangular form and, in addition, transforms the $(k+1)$st column into triangular form. In this way, after at most $n-1$ steps, we arrive at

$$H_{n-1} \cdots H_1 A = R$$

with Householder matrices $H_1, \ldots, H_{n-1}$ and an upper triangular matrix R. From this we obtain

$$A = QR$$

with the unitary matrix

$$Q = (H_{n-1} \cdots H_1)^* = H_1 \cdots H_{n-1}.$$

We summarize our result in the following theorem.

Theorem 2.13 *To each $n \times n$ matrix a QR decomposition can be obtained through $n-1$ Householder transformations.*

The elimination by QR decomposition via Householder matrices can be considered as an alternative to Gaussian elimination, since it does not need pivoting. However, the operation count shows that $2n^3/3 + O(n^2)$ multiplications are required (see Problem 2.18), i.e., twice the cost of Gaussian elimination, and the added expense of partial pivoting in Gaussian elimination does not close this gap. Hence, QR decomposition is rarely used for the solution of linear systems. But later in this book we will see that QR decomposition is an essential part of one of the best algorithms for numerically computing the eigenvalues of a matrix (see Section 7.4).

Problems

2.1 Solve the linear system

$$\begin{aligned} 2x_1 + 4x_2 + x_3 &= 4 \\ 2x_1 + 6x_2 - x_3 &= 10 \\ x_1 + 5x_2 + 2x_3 &= 2 \end{aligned}$$

by Gaussian elimination.

2.2 Write a computer program for the solution of a system of linear equations by Gaussian elimination with partial pivoting and test it for various examples. You will need this code as part of other numerical algorithms later in this book.

2.3 Describe pivoting in Gaussian elimination by using permutation matrices.

2.4 Let A and B be two $n \times n$ matrices. Show that if AB is nonsingular, then A and B are nonsingular.

2.5 Let A, B, C, and D be $n \times n$ matrices and let A be nonsingular. Show that

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det A \det(D - CA^{-1}B).$$

2.6 Verify the summation formulas

$$\sum_{k=1}^{n} k = \frac{1}{2}\, n(n+1) \quad \text{and} \quad \sum_{k=1}^{n} k^2 = \frac{1}{6}\, n(n+1)(2n+1)$$

that were used in the proof of Theorem 2.7.

2.7 Prove the analogue of Theorem 2.7 for the number of additions in Gaussian elimination.

2.8 Show that tridiagonal matrices

$$A = \begin{pmatrix} a_1 & c_1 & & & & \\ b_2 & a_2 & c_2 & & & \\ & b_3 & a_3 & c_3 & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & b_{n-1} & a_{n-1} & c_{n-1} \\ & & & & b_n & a_n \end{pmatrix}$$

with the properties

$$|a_j| \geq |b_j| + |c_j|, \quad b_j c_j \neq 0, \quad j = 2, \ldots, n-1,$$

and $|a_1| > |c_1| > 0$ and $|a_n| > |b_n| > 0$ are nonsingular.

2.9 Show that Gaussian elimination for tridiagonal $n \times n$ matrices requires $4n$ multiplications.

2.10 Show that the diagonal elements of a positive definite matrix are positive.

2.11 Prove that if $A = LL^T$ where L is a real lower triangular nonsingular $n \times n$ matrix, then A is symmetric and positive definite.

2.12 Show that

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 4 \end{pmatrix}$$

is not positive definite.

2.13 Show that for a symmetric positive definite matrix the symmetry and positive definiteness are preserved in Gaussian elimination if diagonal elements are taken as pivot elements, i.e., the submatrices $a_{jk}^{(m)}$, $j, k = m, \dots, n$, are symmetric and positive definite.

2.14 Show that the product of two lower (upper) triangular matrices again is lower (upper) triangular, that lower (upper) triangular matrices with nonvanishing diagonal elements are nonsingular, and that the inverse matrix of a lower (upper) triangular matrix again is lower (upper) triangular.

2.15 Let A be a nonsingular matrix and suppose $A = L_1R_1 = L_2R_2$, where L_1 and L_2 are lower triangular matrices with diagonal elements equal to one and R_1 and R_2 are upper triangular matrices. Show that $L_1 = L_2$ and $R_1 = R_2$.

2.16 Show that for each nonsingular $n \times n$ matrix A there exists a permutation matrix P such that PA has an LR decomposition.

2.17 Solve the linear system

$$\begin{aligned} x_1 + 6x_2 - 2x_3 &= 5 \\ 2x_1 + x_2 - 2x_3 &= 1 \\ 2x_1 + 2x_2 + 6x_3 &= 10 \end{aligned}$$

by QR decomposition.

2.18 Show that the solution of an $n \times n$ linear system by QR elimination with Householder matrices requires $2n^3/3 + O(n^2)$ multiplications.

2.19 Let A be a complex $n \times n$ matrix and $y \in \mathbb{C}^n$ and assume that A, $\operatorname{Re} A$, and $\operatorname{Im} A$ are nonsingular. Show that the $n \times n$ complex linear system $Ax = y$ is equivalent to the two $n \times n$ real systems

$$\{(\operatorname{Im} A)^{-1} \operatorname{Re} A + (\operatorname{Re} A)^{-1} \operatorname{Im} A\} \operatorname{Re} x = (\operatorname{Im} A)^{-1} \operatorname{Re} y + (\operatorname{Re} A)^{-1} \operatorname{Im} y,$$

$$\{(\operatorname{Im} A)^{-1} \operatorname{Re} A + (\operatorname{Re} A)^{-1} \operatorname{Im} A\} \operatorname{Im} x = (\operatorname{Im} A)^{-1} \operatorname{Im} y - (\operatorname{Re} A)^{-1} \operatorname{Re} y.$$

2.20 Use QR decomposition to prove Hadamard's inequality

$$|\det A|^2 \le \prod_{j=1}^{n} \sum_{k=1}^{n} |a_{jk}|^2$$

for the determinant of an $n \times n$ matrix $A = (a_{jk})$.

3

Basic Functional Analysis

In the subsequent chapters we want to discuss iterative methods for the solution of systems of linear and nonlinear equations. For this we will need some fundamental concepts of functional analysis, which we will start to develop now. We shall use these functional-analytic tools also in later parts of this book in some of our convergence and error analysis for the approximate solution of differential and integral equations.

We begin by introducing the notions of normed spaces and their elementary properties, where we assume that the reader is familiar with the concept of linear spaces or vector spaces and their basic properties. Then we proceed by considering scalar product spaces as special cases of normed spaces.

We will continue with the discussion of linear and continuous operators acting between normed spaces. Particular attention is given to linear operators between finite-dimensional spaces, i.e., to matrices and their various norms. The main part of this chapter is Banach's fixed point theorem, also known as the contraction mapping principle, which is one of the most important tools in numerical analysis and is the fundamental basis of our investigations of iterative methods for linear and nonlinear systems. At the end of the chapter we will introduce some of the basic concepts of approximation theory, which will be useful later in other parts of this book.

For a broader and more detailed study we refer to [5, 34, 35, 39, 59] or any other introductory book on functional analysis.

3.1 Normed Spaces

Definition 3.1 *Let X be a complex (or real) linear space (vector space). A function $\|\cdot\| : X \to \mathbb{R}$ with the properties*

$$\begin{array}{llll} \text{(N1)} & \|x\| \geq 0, & & \text{(positivity)} \\ \text{(N2)} & \|x\| = 0 & \textit{if and only if } x = 0, & \text{(definiteness)} \\ \text{(N3)} & \|\alpha x\| = |\alpha|\,\|x\|, & & \text{(homogeneity)} \\ \text{(N4)} & \|x+y\| \leq \|x\| + \|y\|, & & \text{(triangle inequality)} \end{array}$$

for all $x, y \in X$ and all $\alpha \in \mathbb{C}$ (or $\mathbb{R}$) is called a norm *on X. A linear space X equipped with a norm is called a* normed space. *For $X = \mathbb{R}^n$ or $X = \mathbb{C}^n$ we will also call the norm a* vector norm.

Example 3.2 Some examples of norms on $\mathbb{R}^n$ and $\mathbb{C}^n$ are given by

$$\|x\|_1 := \sum_{j=1}^{n} |x_j|, \quad \|x\|_2 := \left(\sum_{j=1}^{n} |x_j|^2 \right)^{1/2}, \quad \|x\|_\infty := \max_{j=1,\ldots,n} |x_j|$$

for $x = (x_1, \ldots, x_n)^T$. It is an easy exercise for the reader to verify that the norm axioms (N1)–(N3) are satisfied. The triangle inequality for the norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ follows immediately from the triangle inequality in $\mathbb{R}$ or $\mathbb{C}$. The verification of the triangle inequality for the norm $\|\cdot\|_2$ is postponed until Section 3.2. □

The norms in Example 3.2 are denoted the ℓ_1, ℓ_2, and ℓ_∞ norm, respectively. For obvious reasons the ℓ_2 norm is also called the *Euclidean norm*, and the ℓ_∞ norm is called the *maximum norm*. The three norms are special cases of the ℓ_p norm

$$\|x\|_p := \left(\sum_{j=1}^{n} |x_j|^p \right)^{1/p}, \tag{3.1}$$

defined for any real number $p \geq 1$. The ℓ_∞ norm is the limiting case of (3.1) as $p \to \infty$ (see Problem 3.1).

Remark 3.3 *For each norm, the* second triangle inequality

$$|\,\|x\| - \|y\|\,| \leq \|x - y\|$$

holds for all $x, y \in X$.

Proof. From the triangle inequality we have

$$\|x\| = \|x - y + y\| \leq \|x - y\| + \|y\|,$$

whence $||x|| - ||y|| \leq ||x - y||$ follows. Analogously, by interchanging the roles of x and y we have $||y|| - ||x|| \leq ||y - x||$. □

For two elements x, y in a normed space $||x - y||$ is called the *distance* between x and y.

Definition 3.4 *A sequence (x_n) of elements in a normed space X is called* convergent *if there exists an element $x \in X$ such that*

$$\lim_{n\to\infty} ||x_n - x|| = 0,$$

i.e., if for every $\varepsilon > 0$ there exists an integer $N(\varepsilon)$ such that $||x_n - x|| < \varepsilon$ for all $n \geq N(\varepsilon)$. The element x is called the limit *of the sequence (x_n), and we write*

$$\lim_{n\to\infty} x_n = x$$

or

$$x_n \to x, \quad n \to \infty.$$

A sequence that does not converge is called divergent.

Theorem 3.5 *The limit of a convergent sequence is uniquely determined.*

Proof. Assume that $x_n \to x$ and $x_n \to y$ for $n \to \infty$. Then from the triangle inequality we obtain that

$$||x - y|| = ||x - x_n + x_n - y|| \leq ||x - x_n|| + ||x_n - y|| \to 0, \quad n \to \infty.$$

Therefore, $||x - y|| = 0$ and $x = y$ by (N2). □

Definition 3.6 *Two norms on a linear space are called* equivalent *if they have the same convergent sequences.*

Theorem 3.7 *Two norms $||\cdot||_a$ and $||\cdot||_b$ on a linear space X are equivalent if and only if there exist positive numbers c and C such that*

$$c||x||_a \leq ||x||_b \leq C||x||_a$$

for all $x \in X$. The limits with respect to the two norms coincide.

Proof. Provided that the conditions are satisfied, from $||x_n - x||_a \to 0$, $n \to \infty$, it follows that $||x_n - x||_b \to 0$, $n \to \infty$, and vice versa.

Conversely, let the two norms be equivalent and assume that there is no $C > 0$ such that $||x||_b \leq C||x||_a$ for all $x \in X$. Then there exists a sequence (x_n) with $||x_n||_a = 1$ and $||x_n||_b \geq n^2$. Now, the sequence (y_n) with $y_n := x_n/n$ converges to zero with respect to $||\cdot||_a$, whereas with respect to $||\cdot||_b$ it is divergent because of $||y_n||_b \geq n$. □

Theorem 3.8 *On a finite-dimensional linear space all norms are equivalent.*

Proof. In a linear space X with finite dimension n and basis $u_1, \dots, u_n$ every element can be expressed in the form

$$x = \sum_{j=1}^{n} \alpha_j u_j.$$

As in Example 3.2,

$$\|x\|_\infty := \max_{j=1,\dots,n} |\alpha_j| \tag{3.2}$$

defines a norm on X. Let $\|\cdot\|$ denote any other norm on X. Then, by the triangle inequality we have

$$\|x\| \le \sum_{j=1}^{n} |\alpha_j| \, \|u_j\| \le C\|x\|_\infty$$

for all $x \in X$, where

$$C := \sum_{j=1}^{n} \|u_j\|.$$

Assume that there is no $c > 0$ such that $c\|x\|_\infty \le \|x\|$ for all $x \in X$. Then there exists a sequence (x_ν) with $\|x_\nu\| = 1$ such that $\|x_\nu\|_\infty \ge \nu$. Consider the sequence (y_ν) with $y_\nu := x_\nu / \|x_\nu\|_\infty$ and write

$$y_\nu = \sum_{j=1}^{n} \alpha_{j\nu} u_j.$$

Because of $\|y_\nu\|_\infty = 1$ each of the sequences $(\alpha_{j\nu})$, $j = 1, \dots, n$, is bounded in $\mathbb{C}$. Hence, by the Bolzano–Weierstrass theorem we can select convergent subsequences $\alpha_{j,\nu(\ell)} \to \alpha_j$, $\ell \to \infty$, for each $j = 1, \dots, n$. This now implies $\|y_{\nu(\ell)} - y\|_\infty \to 0$, $\ell \to \infty$, where

$$y := \sum_{j=1}^{n} \alpha_j u_j,$$

and also $\|y_{\nu(\ell)} - y\| \le C\|y_{\nu(\ell)} - y\|_\infty \to 0$, $\ell \to \infty$. But on the other hand we have $\|y_\nu\| = 1/\|x_\nu\|_\infty \to 0$, $\nu \to \infty$. Therefore, $y = 0$, and consequently $\|y_{\nu(\ell)}\|_\infty \to 0$, $\ell \to \infty$, which contradicts $\|y_\nu\|_\infty = 1$ for all ν. □

The following definitions carry over some useful concepts from Euclidean space to general normed spaces.

Definition 3.9 *A subset U of a normed space X is called* closed *if it contains all limits of convergent sequences of U. The* closure $\overline{U}$ *of a subset U of a normed space X is the set of all limits of convergent sequences of U. A subset U is called* open *if its complement $X \setminus U$ is closed. A set U is called* dense *in another set V if $V \subset \overline{U}$, i.e., if each element in V is the limit of a convergent sequence from U.*

Obviously, a subset U is closed if and only if it coincides with its closure. For x_0 in X and $r > 0$ the set $B[x_0, r] := \{x \in X : \|x - x_0\| \le r\}$ is closed and is called the *closed ball* of radius r and center x_0. Correspondingly, the set $B(x_0, r) := \{x \in X : \|x - x_0\| < r\}$ is open and is called an *open ball.*

Definition 3.10 *A subset U of a normed space X is called* bounded *if there exists a positive number C such that $\|x\| \le C$ for all $x \in U$.*

Convergent sequences are bounded (see Problem 3.6).

Theorem 3.11 *Any bounded sequence in a finite-dimensional normed space X contains a convergent subsequence.*

Proof. Let $u_1, \dots, u_n$ be a basis of X and let (x_ν) be a bounded sequence. Then writing

$$x_\nu = \sum_{j=1}^{n} \alpha_{j\nu} u_j$$

and using the norm (3.2), as in the proof of Theorem 3.8 we deduce that each of the sequences $(\alpha_{j\nu})$, $j = 1, \dots, n$, is bounded in $\mathbb{C}$. Hence, by the Bolzano–Weierstrass theorem we can select convergent subsequences $\alpha_{j,\nu(\ell)} \to \alpha_j$, $\ell \to \infty$, for each $j = 1, \dots, n$. This now implies

$$x_{\nu(\ell)} \to \sum_{j=1}^{n} \alpha_j u_j \in X, \quad \ell \to \infty,$$

and the proof is finished. □

3.2 Scalar Products

Definition 3.12 *Let X be a complex (or real) linear space. Then a function $(\cdot\,,\cdot) : X \times X \to \mathbb{C}$ (or $\mathbb{R}$) with the properties*

(H1) $\quad (x, x) \ge 0,$ (positivity)

(H2) $\quad (x, x) = 0$ *if and only if* $x = 0,$ (definiteness)

(H3) $\quad (x, y) = \overline{(y, x)},$ (symmetry)

(H4) $\quad (\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z),$ (linearity)

for all $x, y, z \in X$ and $\alpha, \beta \in \mathbb{C}$ (or $\mathbb{R}$) is called a scalar product, *or an* inner product, *on X. (By the bar we denote the complex conjugate.) A linear space X equipped with a scalar product is called a* pre-Hilbert space.

As a simple consequence of (H3) and (H4) we note the antilinearity

$$\text{(H4')} \qquad (x, \alpha y + \beta z) = \bar{\alpha}(x, y) + \bar{\beta}(x, z).$$

Example 3.13 *An example of a scalar product on* $\mathbb{R}^n$ *and* $\mathbb{C}^n$ *is given by*

$$(x, y) := \sum_{j=1}^{n} x_i \bar{y}_i$$

for $x = (x_1, \ldots, x_n)^T$ *and* $y = (y_1, \ldots, y_n)^T$. *(Note that* $(x, y) = y^* x$.*)*

Theorem 3.14 *For a scalar product we have the* Cauchy–Schwarz inequality

$$|(x, y)|^2 \leq (x, x)(y, y)$$

for all $x, y \in X$, *with equality if and only if* x *and* y *are linearly dependent.*

Proof. The inequality is trivial for $x = 0$. For $x \neq 0$ it follows from

$$\begin{aligned}(\alpha x + \beta y, \alpha x + \beta y) &= |\alpha|^2 (x, x) + 2\,\mathrm{Re}\{\alpha \bar{\beta}(x, y)\} + |\beta|^2 (y, y) \\ &= (x, x)(y, y) - |(x, y)|^2,\end{aligned}$$

where we have set $\alpha = -(x, x)^{-1/2}\overline{(x, y)}$ and $\beta = (x, x)^{1/2}$. Since $(\cdot\,, \cdot)$ is positive definite, this expression is nonnegative, and it is equal to zero if and only if $\alpha x + \beta y = 0$. In the latter case x and y are linearly dependent because $\beta \neq 0$. □

Theorem 3.15 *A scalar product* $(\cdot\,, \cdot)$ *on a linear space* X *defines a norm by*

$$\|x\| := (x, x)^{1/2}$$

for all $x \in X$; *i.e., a pre-Hilbert space is always a normed space.*

Proof. We leave it as an exercise for the reader to verify the norm axioms. The triangle inequality follows by

$$\|x + y\|^2 = (x + y, x + y) \leq \|x\|^2 + 2\|x\|\,\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2$$

from the Cauchy–Schwarz inequality. □

Note that we can rewrite the Cauchy–Schwarz inequality in the form

$$|(x, y)| \leq \|x\|\,\|y\|.$$

The scalar product of Example 3.13 generates the Euclidean norm of Example 3.2, and therefore it is called the Euclidean scalar product. Theorem 3.15 includes the triangle inequality for the Euclidean norm that we postponed in Example 3.2.

The following definition generalizes the concept of orthogonality from Euclidean space to pre-Hilbert spaces.

Definition 3.16 *Two elements x and y of a pre-Hilbert space X are called* orthogonal *if*

$$(x, y) = 0.$$

Two subsets U and V of X are called orthogonal if each pair of elements $x \in U$ and $y \in V$ are orthogonal. For two orthogonal elements or subsets we write $x \perp y$ and $U \perp V$, respectively. A subset U of X is called an orthogonal system *if $(x, y) = 0$ for all $x, y \in U$ with $x \neq y$. An orthogonal system U is called an* orthonormal system *if $\|x\| = 1$ for all $x \in U$.*

Theorem 3.17 *The elements of an orthonormal system are linearly independent.*

Proof. From

$$\sum_{k=1}^{n} \alpha_k q_k = 0$$

for the orthonormal system $\{q_1, \dots, q_n\}$, by taking the scalar product with q_j, we immediately have that $\alpha_j = 0$ for $j = 1, \dots, n$. □

The *Gram–Schmidt orthogonalization* procedure as described in the following theorem provides a converse of Theorem 3.17. For a subset U of a linear space X we denote the set *spanned* by all linear combinations of elements of U by $\operatorname{span}\{U\}$.

Theorem 3.18 *Let $\{u_0, u_1, \dots\}$ be a finite or countable number of linearly independent elements of a pre-Hilbert space. Then there exists a uniquely determined orthogonal system $\{q_0, q_1, \dots\}$ of the form*

$$q_n = u_n + r_n, \quad n = 0, 1, \dots, \tag{3.3}$$

with $r_0 = 0$ and $r_n \in \operatorname{span}\{u_0, \dots, u_{n-1}\}$, $n = 1, 2, \dots,$ satisfying

$$\operatorname{span}\{u_0, \dots, u_n\} = \operatorname{span}\{q_0, \dots, q_n\}, \quad n = 0, 1, \dots. \tag{3.4}$$

Proof. Assume that we have constructed orthogonal elements of the form (3.3) with the property (3.4) up to q_{n-1}. By (3.4), the $\{q_0, \dots, q_{n-1}\}$ are linearly independent, and therefore $\|q_k\| \neq 0$ for $k = 0, 1, \dots, n-1$. Hence,

$$q_n := u_n - \sum_{k=0}^{n-1} \frac{(u_n, q_k)}{(q_k, q_k)} \, q_k$$

is well-defined, and using the induction assumption, we obtain $(q_n, q_m) = 0$ for $m = 0, \dots, n-1$ and

$$\operatorname{span}\{u_0, \dots, u_{n-1}, u_n\} = \operatorname{span}\{q_0, \dots, q_{n-1}, u_n\} = \operatorname{span}\{q_0, \dots, q_{n-1}, q_n\}.$$

Hence, the existence of q_n is established.

Assume that $\{q_0, q_1, \dots\}$ and $\{\tilde{q}_0, \tilde{q}_1, \dots\}$ are two orthogonal sets of elements with the required properties. Then clearly $q_0 = u_0 = \tilde{q}_0$. Assume that we have shown that equality holds up to $q_{n-1} = \tilde{q}_{n-1}$. Then, since $q_n - \tilde{q}_n \in \operatorname{span}\{u_0, \dots, u_{n-1}\}$, we can represent $q_n - \tilde{q}_n$ as a linear combination of $q_1, \dots, q_{n-1}$; i.e.,

$$q_n - \tilde{q}_n = \sum_{k=0}^{n-1} \alpha_k q_k.$$

Now the orthogonality yields

$$\|q_n - \tilde{q}_n\|^2 = \left(q_n - \tilde{q}_n, \sum_{k=0}^{n-1} \alpha_k q_k \right) = 0,$$

whence $q_n = \tilde{q}_n$. □

3.3 Bounded Linear Operators

By the symbol $A : X \to Y$ we will denote a mapping whose domain of definition is a set X and whose range is contained in a set Y; i.e., for every $x \in X$ the mapping A assigns a unique element $Ax \in Y$. The *range* is the set $A(X) := \{Ax : x \in X\}$ of all image elements. We will use the terms *mapping, function,* and *operator* synonymously. (We have already used this convention in Definitions 3.1 and 3.12.)

Definition 3.19 *An operator A mapping a subset U of a normed space X into a normed space Y is called* continuous *at $x \in U$ if for every sequence (x_n) from U with $\lim_{n\to\infty} x_n = x$ we have $\lim_{n\to\infty} Ax_n = Ax$. The function $A : U \to Y$ is called continuous if it is continuous for all $x \in U$.*

An equivalent definition is the following: A function $A : U \subset X \to Y$ is continuous at $x \in U$ if for every $\varepsilon > 0$ there exists $\delta > 0$ such that $\|Ax - Ay\| < \varepsilon$ for all $y \in U$ with $\|x - y\| < \delta$. Here we have used the same symbol $\|\cdot\|$ for the norms on X and Y. Note that by the second triangle inequality of Remark 3.3 the norm is a continuous function.

Definition 3.20 *An operator $A : X \to Y$ mapping a linear space X into a linear space Y is called* linear *if*

$$A(\alpha x + \beta y) = \alpha Ax + \beta Ay$$

for all $x, y \in X$ and all $\alpha, \beta \in \mathbb{C}$ (or $\mathbb{R}$).

Theorem 3.21 *A linear operator is continuous if it is continuous at one element.*

Proof. Let $A : X \to Y$ be continuous at $x_0 \in X$. Then for every $x \in X$ and every sequence (x_n) with $x_n \to x$, $n \to \infty$, we have

$$Ax_n = A(x_n - x + x_0) + A(x - x_0) \to A(x_0) + A(x - x_0) = A(x), \quad n \to \infty,$$

since $x_n - x + x_0 \to x_0$, $n \to \infty$. □

Definition 3.22 *A linear operator $A : X \to Y$ from a normed space X into a normed space Y is called* bounded *if there exists a positive number C such that*

$$\|Ax\| \le C\|x\|$$

for all $x \in X$. Each number C for which this inequality holds is called a bound *for the operator A. (Again we have used the same symbol $\|\cdot\|$ for the norms on X and Y.)*

Theorem 3.23 *A linear operator $A : X \to Y$ is bounded if and only if*

$$\|A\| := \sup_{\|x\|=1} \|Ax\| < \infty.$$

The number $\|A\|$ is the smallest bound for A and is called the norm *of A.*

Proof. Assume that A is bounded with the bound C. Then

$$\sup_{\|x\|=1} \|Ax\| \le C,$$

and, in particular, $\|A\|$ is less than or equal to any bound for A. Conversely, if $\|A\| < \infty$, then using the linearity of A and the homogeneity of the norm, we find that

$$\|Ax\| = \left\| A\left(\frac{x}{\|x\|}\right)\right\| \|x\| \le \|A\| \, \|x\|$$

for all $x \neq 0$. Therefore, A is bounded with the bound $\|A\|$. □

Theorem 3.24 *A linear operator is continuous if and only if it is bounded.*

Proof. Let $A : X \to Y$ be bounded and let (x_n) be a sequence in X with $x_n \to 0$, $n \to \infty$. Then from $\|Ax_n\| \le C\|x_n\|$ it follows that $Ax_n \to 0$, $n \to \infty$. Thus, A is continuous at $x = 0$, and because of Theorem 3.21 it is continuous everywhere in X.

Conversely, let A be continuous and assume that there is no $C > 0$ such that $\|Ax\| \le C\|x\|$ for all $x \in X$. Then there exists a sequence (x_n) in X with $\|x_n\| = 1$ and $\|Ax_n\| \ge n$. Consider the sequence $y_n := x_n/\|Ax_n\|$. Then $y_n \to 0$, $n \to \infty$, and since A is continuous, $Ay_n \to A(0) = 0$, $n \to \infty$. This is a contradiction to $\|Ay_n\| = 1$ for all n. Hence, A is bounded. □

Remark 3.25 *Let X, Y, and Z be normed spaces and let $A : X \to Y$ and $B : Y \to Z$ be bounded linear operators. Then the product $BA : X \to Z$, defined by $(BA)x := B(Ax)$ for all $x \in X$, is a bounded linear operator with $\|BA\| \le \|A\| \, \|B\|$.*

Proof. This follows from $\|(BA)x\| = \|B(Ax)\| \le \|B\| \, \|A\| \, \|x\|$. □

3.4 Matrix Norms

Theorem 3.26 *Let (a_{jk}) be a real or complex $n \times n$ matrix. Then the linear operators $A : \mathbb{R}^n \to \mathbb{R}^n$ and $A : \mathbb{C}^n \to \mathbb{C}^n$, defined by*

$$(Ax)_j := \sum_{k=1}^{n} a_{jk} x_k, \quad j = 1, \dots, n,$$

are bounded with respect to each norm on $\mathbb{R}^n$ and $\mathbb{C}^n$. In particular, we have

$$\|A\|_1 = \max_{k=1,\dots,n} \sum_{j=1}^{n} |a_{jk}|, \tag{3.5}$$

$$\|A\|_\infty = \max_{j=1,\dots,n} \sum_{k=1}^{n} |a_{jk}|, \tag{3.6}$$

$$\|A\|_2 \le \left(\sum_{j,k=1}^{n} |a_{jk}|^2 \right)^{1/2}. \tag{3.7}$$

In this case the norms are also called matrix norms. *(Note that in (3.5)–(3.7) both the domain and the range are given the same norm.)*

Proof. By Theorem 3.8 it suffices to prove boundedness of A with respect to one norm. For $\|\cdot\|_1$ we can estimate

$$\|Ax\|_1 = \sum_{j=1}^{n} |(Ax)_j| = \sum_{j=1}^{n} \left| \sum_{k=1}^{n} a_{jk} x_k \right|$$

$$\le \sum_{k=1}^{n} |x_k| \sum_{j=1}^{n} |a_{jk}| \le \max_{k=1,\dots,n} \sum_{j=1}^{n} |a_{jk}| \sum_{k=1}^{n} |x_k|.$$

Therefore, we have that

$$\|A\|_1 \le \max_{k=1,\dots,n} \sum_{j=1}^{n} |a_{jk}|. \tag{3.8}$$

Now choose i such that

$$\sum_{j=1}^{n} |a_{ji}| = \max_{k=1,\dots,n} \sum_{j=1}^{n} |a_{jk}|,$$

and choose $z \in \mathbb{R}^n$ with $z_i = 1$ and $z_k = 0$ for $k \neq i$. Then $\|z\|_1 = 1$ and

$$\|Az\|_1 = \sum_{j=1}^{n} |(Az)_j| = \sum_{j=1}^{n} \left| \sum_{k=1}^{n} a_{jk} z_k \right| = \sum_{j=1}^{n} |a_{ji}| = \max_{k=1,\dots,n} \sum_{j=1}^{n} |a_{jk}|.$$

Hence

$$\|A\|_1 = \sup_{\|x\|_1=1} \|Ax\|_1 \geq \|Az\|_1 = \max_{k=1,\dots,n} \sum_{j=1}^{n} |a_{jk}|, \tag{3.9}$$

and from (3.8) and (3.9) we obtain (3.5).

For $\|\cdot\|_\infty$ we can estimate

$$\|Ax\|_\infty = \max_{j=1,\dots,n} |(Ax)_j| = \max_{j=1,\dots,n} \left| \sum_{k=1}^{n} a_{jk} x_k \right|$$

$$\leq \max_{j=1,\dots,n} \sum_{k=1}^{n} |a_{jk}|\,|x_k| \leq \max_{j=1,\dots,n} \sum_{k=1}^{n} |a_{jk}| \max_{k=1,\dots,n} |x_k|.$$

Therefore, we have that

$$\|A\|_\infty \leq \max_{j=1,\dots,n} \sum_{k=1}^{n} |a_{jk}|. \tag{3.10}$$

Now choose i such that

$$\sum_{k=1}^{n} |a_{ik}| = \max_{j=1,\dots,n} \sum_{k=1}^{n} |a_{jk}|,$$

and choose $z \in \mathbb{C}^n$ with $z_k = \bar{a}_{ik}/|a_{ik}|$ if $a_{ik} \neq 0$ and $z_k = 1$ if $a_{ik} = 0$. Then $\|z\|_\infty = 1$ and

$$\|Az\|_\infty = \max_{j=1,\dots,n} |(Az)_j| = \max_{j=1,\dots,n} \left| \sum_{k=1}^{n} a_{jk} z_k \right|$$

$$\geq \left| \sum_{k=1}^{n} a_{ik} z_k \right| = \sum_{k=1}^{n} |a_{ik}| = \max_{j=1,\dots,n} \sum_{k=1}^{n} |a_{jk}|.$$

Hence

$$\|A\|_\infty = \sup_{\|x\|_\infty=1} \|Ax\|_\infty \geq \|Az\|_\infty = \max_{j=1,\dots,n} \sum_{k=1}^{n} |a_{jk}|, \tag{3.11}$$

and from (3.10) and (3.11) we obtain (3.6).

Finally, for $\|\cdot\|_2$, using the Cauchy–Schwarz inequality we can estimate

$$\|Ax\|_2^2 = \sum_{j=1}^{n} |(Ax)_j|^2 = \sum_{j=1}^{n} \left| \sum_{k=1}^{n} a_{jk} x_k \right|^2$$

$$\leq \sum_{j=1}^{n} \left\{ \sum_{k=1}^{n} |a_{jk}|^2 \sum_{k=1}^{n} |x_k|^2 \right\} = \sum_{j,k=1}^{n} |a_{jk}|^2 \sum_{k=1}^{n} |x_k|^2.$$

Therefore,

$$\|A\|_2^2 \leq \sum_{j,k=1}^{n} |a_{jk}|^2,$$

and (3.7) is proven. In this inequality equality does not hold, in general, as can be seen by considering the identity matrix. □

In order to derive a representation for $\|A\|_2$ we need to recall the definition and some basic facts about eigenvalues and eigenvectors of a matrix. A number $\lambda \in \mathbb{C}$ is called an *eigenvalue* of the matrix A if there exists a vector $x \in \mathbb{C}^n$ with $x \neq 0$ such that

$$Ax = \lambda x.$$

The vector x is called an *eigenvector* for the eigenvalue λ. Each $n \times n$ matrix has at least one and at most n eigenvalues, since the *characteristic polynomial* $\det(A - \lambda I)$ has at least one and at most n zeros. Eigenvectors for different eigenvalues are linearly independent (see Problem 3.12). The *algebraic multiplicity* of an eigenvalue of a matrix is its multiplicity as a zero of the characteristic polynomial; its *geometric multiplicity* is the number of linearly independent eigenvectors associated with the eigenvalue.

Theorem 3.27 *To each matrix A there exists a unitary matrix Q such that Q^*AQ is an upper triangular matrix.*

Proof. Assume that it has been shown that for each $(n-1) \times (n-1)$ matrix A_{n-1} there exists a unitary $(n-1) \times (n-1)$ matrix Q_{n-1} such that $Q_{n-1}^* A_{n-1} Q_{n-1}$ is an upper triangular matrix. Let λ be an eigenvalue of the $n \times n$ matrix A_n with eigenvector u. We may assume that $(u, u) = 1$, where $(\cdot\,,\cdot)$ is the Euclidean scalar product. Using the Gram–Schmidt procedure of Theorem 3.18 we can construct an orthonormal basis of $\mathbb{C}^n$ of the form $u, v_2, \ldots, v_n$. Then we define a unitary $n \times n$ matrix by

$$U_n := (u, v_2, \ldots, v_n).$$

With the aid of $(u, v_j) = 0$, $j = 2, \ldots, n$, we see that

$$U_n^* A_n U_n = U_n^*(\lambda u, A_n v_2, \ldots, A_n v_n) = \begin{pmatrix} \lambda & * \\ 0 & A_{n-1} \end{pmatrix},$$

with some $(n-1) \times (n-1)$ matrix A_{n-1}. By the induction assumption there exists a unitary $(n-1) \times (n-1)$ matrix Q_{n-1} such that $Q_{n-1}^* A_{n-1} Q_{n-1}$ is upper triangular. Then

$$Q_n := U_n \begin{pmatrix} 1 & 0 \\ 0 & Q_{n-1} \end{pmatrix}$$

defines a unitary $n \times n$ matrix, and $Q_n^* A_n Q_n$ is upper triangular. □

Lemma 3.28 *For an $n \times n$ matrix A and its adjoint A^* we have that*

$$(Ax, y) = (x, A^*y)$$

for all $x, y \in \mathbb{C}^n$, where $(\cdot\,,\cdot)$ denotes the Euclidean scalar product.

Proof. Simple calculations yield

$$\begin{aligned}(Ax, y) &= \sum_{j=1}^{n} (Ax)_j \bar{y}_j = \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk} x_k \bar{y}_j \\ &= \sum_{k=1}^{n}\sum_{j=1}^{n} x_k \overline{a_{kj}^* y_j} = \sum_{k=1}^{n} x_k \overline{A^* y_k} = (x, A^*y),\end{aligned}$$

where we have used that $a_{kj}^* = \overline{a_{jk}}$. □

Theorem 3.29 *The eigenvalues of a* Hermitian *$n \times n$ matrix are real, and the eigenvectors form an orthogonal basis in $\mathbb{C}^n$.*

Proof. If A is Hermitian, i.e., if $A = A^*$, then the matrix $\tilde{A} := Q^*AQ$ from Theorem 3.27 is also Hermitian, since

$$\tilde{A}^* = (Q^*AQ)^* = Q^*A^*Q^{**} = Q^*AQ = \tilde{A}.$$

Therefore, in this case the upper triangular matrix $\tilde{A}$ must be diagonal; i.e.,

$$\tilde{A} = D := \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

Since from $Q^*AQ = D$ it follows that $AQ = QD$, we can conclude that the columns of $Q = (u_1, \ldots, u_n)$ satisfy $Au_j = \lambda_j u_j$, $j = 1, \ldots, n$. Hence the eigenvectors of a Hermitian matrix form an orthogonal basis in $\mathbb{C}^n$. Because of

$$\lambda_j = (Au_j, u_j) = (u_j, Au_j) = \overline{(Au_j, u_j)} = \bar{\lambda}_j,$$

the eigenvalues of Hermitian matrices are real. □

For a *positive semidefinite* matrix A, i.e., for a Hermitian matrix with the property

$$(Ax, x) \geq 0, \quad x \in \mathbb{C}^n,$$

all eigenvalues are real and nonnegative. Analogously, the eigenvalues of a *positive definite* matrix A, i.e., of a Hermitian matrix with the property

$$(Ax, x) > 0, \quad x \in \mathbb{C}^n,\ x \neq 0,$$

are positive.

Definition 3.30 *The number*

$$\rho(A) := \max\,\{|\lambda| : \lambda \text{ eigenvalue of } A\}$$

is called the spectral radius *of* A.

Theorem 3.31 *For an* $n \times n$ *matrix* A *we have*

$$\|A\|_2 = \sqrt{\rho(A^*A)}.$$

If A *is Hermitian, then*

$$\|A\|_2 = \rho(A).$$

Proof. From Lemma 3.28 we have that

$$\|Ax\|_2^2 = (Ax, Ax) = (x, A^*Ax)$$

for all $x \in \mathbb{C}^n$. Hence the Hermitian matrix A^*A is positive semidefinite and therefore has n orthonormal eigenvectors

$$A^*Au_j = \mu_j^2 u_j, \quad j = 1, \ldots, n,$$

with real nonnegative eigenvalues. We use the orthonormal basis of eigenvectors and represent $x \in \mathbb{C}^n$ by

$$x = \sum_{j=1}^n \alpha_j u_j$$

and have

$$\|x\|_2^2 = (x, x) = \left(\sum_{j=1}^n \alpha_j u_j, \sum_{k=1}^n \alpha_k u_k\right) = \sum_{j=1}^n |\alpha_j|^2$$

and

$$\|Ax\|_2^2 = (Ax, Ax) = (x, A^*Ax) = \left(\sum_{j=1}^n \alpha_j u_j, \sum_{k=1}^n \mu_k^2 \alpha_k u_k\right) = \sum_{j=1}^n \mu_j^2 |\alpha_j|^2.$$

From this we obtain that

$$\|Ax\|_2^2 \le \rho(A^*A)\|x\|_2^2,$$

whence

$$\|A\|_2^2 \le \rho(A^*A)$$

follows. On the other hand, if we choose j such that $\mu_j^2 = \rho(A^*A)$, then we have that

$$\|A\|_2^2 = [\sup_{\|x\|_2=1} \|Ax\|_2]^2 \ge \|Au_j\|_2^2 = (u_j, A^*Au_j) = \mu_j^2 = \rho(A^*A).$$

This concludes the proof of $\|A\|_2 = \sqrt{\rho(A^*A)}$. If A is Hermitian, then $A^*A = A^2$, whence $\rho(A^*A) = \rho(A^2) = [\rho(A)]^2$ follows. □

The following final theorem of this section is of basic importance for establishing a necessary and sufficient condition for the convergence of iterative methods for linear systems.

Theorem 3.32 *For each norm on $\mathbb{C}^n$ and each $n \times n$ matrix A we have that*

$$\rho(A) \le \|A\|.$$

Conversely, to each matrix A and each $\varepsilon > 0$ there exists a norm on $\mathbb{C}^n$ such that

$$\|A\| \le \rho(A) + \varepsilon.$$

Proof. Let λ be an eigenvalue of A with eigenvector u. We may assume that $\|u\| = 1$. Then the first part of the theorem follows from

$$\|A\| = \sup_{\|x\|=1} \|Ax\| \ge \|Au\| = \|\lambda u\| = |\lambda|.$$

For the second part, by Theorem 3.27 there exists a unitary matrix Q such that

$$B = Q^*AQ = \begin{pmatrix} b_{11} & b_{12} & b_{13} & \cdot & \cdot & b_{1n} \\ & b_{22} & b_{23} & \cdot & \cdot & b_{2n} \\ & & b_{33} & \cdot & \cdot & b_{3n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & & b_{nn} \end{pmatrix}$$

is upper triangular. Because of $\det(\lambda I - A) = \det(\lambda I - B)$, the eigenvalues of A are given by $\lambda_j = b_{jj}$, $j = 1, \dots, n$. We set

$$b := \max_{1 \le j \le k \le n} |b_{jk}|$$

and

$$\delta := \min\left(1, \frac{\varepsilon}{(n-1)b}\right)$$

and define the diagonal matrix

$$D := \operatorname{diag}(1, \delta, \delta^2, \dots, \delta^{n-1})$$

with the inverse

$$D^{-1} = \operatorname{diag}(1, \delta^{-1}, \delta^{-2}, \dots, \delta^{-n+1}).$$

Then for $C := D^{-1}BD$ we have that

$$C = \begin{pmatrix} b_{11} & \delta b_{12} & \delta^2 b_{13} & \cdot & \cdot & \delta^{n-1} b_{1n} \\ & b_{22} & \delta b_{23} & \cdot & \cdot & \delta^{n-2} b_{2n} \\ & & b_{33} & \cdot & \cdot & \delta^{n-3} b_{3n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & & b_{nn} \end{pmatrix}.$$

Since $\delta \leq 1$, by Theorem 3.26, we can estimate

$$\|C\|_\infty \leq \max_{j=1,\dots,n} |b_{jj}| + (n-1)\delta b \leq \rho(A) + \varepsilon.$$

After setting $V := QD$ we define a norm on $\mathbb{C}^n$ by $\|x\| := \|V^{-1}x\|_\infty$. Using $C = V^{-1}AV$ we now obtain

$$\|Ax\| = \|V^{-1}Ax\|_\infty = \|CV^{-1}x\|_\infty \leq \|C\|_\infty \|V^{-1}x\|_\infty = \|C\|_\infty \|x\|$$

for all $x \in \mathbb{C}^n$. Hence

$$\|A\| \leq \|C\|_\infty \leq \rho(A) + \varepsilon,$$

and the proof is finished. □

3.5 Completeness

Definition 3.33 *A sequence* (x_n) *of elements in a normed space* X *is called a* Cauchy sequence *if for every* $\varepsilon > 0$ *there exists an integer* $N(\varepsilon)$ *such that*

$$\|x_n - x_m\| < \varepsilon$$

for all $n, m \geq N(\varepsilon)$, *i.e., if* $\lim_{n,m\to\infty} \|x_n - x_m\| = 0$.

Theorem 3.34 *Every convergent sequence is a Cauchy sequence.*

Proof. Let $x_n \to x$, $n \to \infty$. Then, for $\varepsilon > 0$ there exists $N(\varepsilon) \in \mathbb{N}$ such that $\|x_n - x\| < \varepsilon/2$ for all $n \geq N(\varepsilon)$. Now the triangle inequality yields

$$\|x_n - x_m\| = \|x_n - x + x - x_m\| \leq \|x_n - x\| + \|x - x_m\| < \varepsilon$$

for all $n, m \geq N(\varepsilon)$. □

The fact that the converse of Theorem 3.34 is not true in general gives rise to the following definition.

Definition 3.35 *A subset* U *of a normed space* X *is called* complete *if every Cauchy sequence of elements in* U *converges to an element in* U. *A normed space is called a* Banach space *if it is complete. A pre-Hilbert space is called a* Hilbert space *if it is complete.*

The subset of rational numbers is not complete in $\mathbb{R}$. In order to give further examples, we introduce some infinite-dimensional normed spaces.

The set $C[a, b]$ of continuous functions $f : [a, b] \to \mathbb{R}$ equipped with pointwise addition and scalar multiplication,

$$(f + g)(x) := f(x) + g(x), \quad (\alpha f)(x) := \alpha f(x),$$

obviously is a linear space. Since the monomials $x \mapsto x^n$, $n = 0, 1, \dots$, are linearly independent (see Theorem 8.2), $C[a, b]$ has infinite dimension.

Example 3.36 *The linear space $C[a,b]$ furnished with the* maximum norm

$$\|f\|_\infty := \max_{x\in[a,b]} |f(x)|$$

is a Banach space.

Proof. The norm axioms (N1)–(N3) are trivially satisfied. The triangle inequality follows from

$$\|f+g\|_\infty = \max_{x\in[a,b]} |(f+g)(x)| = |(f+g)(x_0)| \le |f(x_0)| + |g(x_0)|$$

$$\le \max_{x\in[a,b]} |f(x)| + \max_{x\in[a,b]} |g(x)| = \|f\|_\infty + \|g\|_\infty$$

for some $x_0 \in [a,b]$. Since the condition $\|f_n - f\|_\infty < \varepsilon$ is equivalent to $|f_n(x) - f(x)| < \varepsilon$ for all $x \in [a,b]$, convergence of a sequence of continuous functions in the maximum norm is equivalent to uniform convergence on $[a,b]$. Since the Cauchy criterion is sufficient for uniform convergence of a sequence of continuous functions to a continuous limit function, the space $C[a,b]$ is complete with respect to the maximum norm. □

Example 3.37 *The linear space $C[a,b]$ equipped with the L_1* norm

$$\|f\|_1 := \int_a^b |f(x)|\,dx$$

is not complete.

Proof. The norm axioms are trivially satisfied. Without loss of generality we take $[a,b] = [0,2]$ and choose

$$f_n(x) := \begin{cases} x^n, & 0 \le x \le 1, \\ 1, & 1 \le x \le 2. \end{cases}$$

Then for $m > n$ we have that

$$\|f_n - f_m\|_1 = \int_0^1 (x^n - x^m)\,dx \le \frac{1}{n+1} \to 0, \quad n \to \infty,$$

and therefore (f_n) is a Cauchy sequence. Now we assume that (f_n) converges with respect to the L_1 norm to a continuous function f; i.e.,

$$\|f_n - f\|_1 \to 0, \quad n \to \infty.$$

Then

$$\int_0^1 |f(x)|\,dx \le \int_0^1 |f(x) - x^n|\,dx + \int_0^1 x^n\,dx \le \|f - f_n\|_1 + \frac{1}{n+1} \to 0$$

for $n \to \infty$, whence $f(x) = 0$ follows for $0 \le x \le 1$. Furthermore, we have

$$\int_1^2 |f(x) - 1|\, dx = \int_1^2 |f(x) - f_n(x)|\, dx \le \|f - f_n\|_1 \to 0, \quad n \to \infty.$$

This implies that $f(x) = 1$ for $1 \le x \le 2$, and we have a contradiction, since f is continuous.

However, we note that the space $L^1[a, b]$ of measurable and Lebesgue integrable real-valued functions is complete with respect to the L_1 norm (see [5, 51, 59]). □

Example 3.38 *The linear space $C[a, b]$ equipped with the L_2* norm

$$\|f\|_2 := \left(\int_a^b |f(x)|^2 dx \right)^{1/2}$$

is not complete.

Proof. The norm is generated by the scalar product

$$(f, g) := \int_a^b f(x) g(x)\, dx.$$

Considering the same sequence as in Example 3.37, it can be seen that $C[a, b]$ also is not complete with respect to the L_2 norm. Again note that the space $L^2[a, b]$ of measurable and Lebesgue square-integrable real-valued functions is complete with respect to the L_2 norm (see [5, 51, 59]). □

Theorem 3.39 *Each finite-dimensional normed space is a Banach space.*

Proof. Let X be finite-dimensional with basis $u_1, \dots, u_n$ and assume that (x_ν) is a Cauchy sequence in X. We represent

$$x_\nu = \sum_{j=1}^n \alpha_{j\nu} u_j$$

and recall from Theorem 3.8 that there exists $C > 0$ such that

$$\max_{j=1,\dots,n} |\alpha_{j\nu} - \alpha_{j\mu}| \le C \|x_\nu - x_\mu\|$$

for all $\nu, \mu \in \mathbb{N}$. Hence for $j = 1, \dots, n$ the $(\alpha_{j\nu})$ are Cauchy sequences in $\mathbb{C}$. Therefore, there exist $\alpha_1, \dots, \alpha_n$ such that $\alpha_{j\nu} \to \alpha_j$, $\nu \to \infty$, for $j = 1, \dots, n$, since the Cauchy criterion is sufficient for convergence in $\mathbb{C}$. Then we have convergence,

$$x_\nu \to x := \sum_{j=1}^n \alpha_j u_j \in X, \quad \nu \to \infty,$$

and the proof is finished. □

Remark 3.40 *Complete sets are closed, and each closed subset of a complete subset is complete.*

Proof. This is trivial. □

3.6 The Banach Fixed Point Theorem

Definition 3.41 *Let U be a subset of a normed space X. An operator $A : U \to X$ is called a* contraction operator *if there exists a constant $q \in [0, 1)$ such that*

$$||Ax - Ay|| \leq q||x - y||$$

for all $x, y \in U$. Each constant q satisfying this inequality is called a contraction number *of the operator A.*

Frequently, we will call a contraction operator simply a contraction.

Remark 3.42 *Each contraction operator is continuous.*

Proof. This is trivial, since the convergence $||x_n - x|| \to 0$, $n \to \infty$, implies that $||Ax_n - Ax|| \leq q||x_n - x|| \to 0$, $n \to \infty$. □

An operator $A : U \to X$ is called *Lipschitz continuous* with Lipschitz constant L if there exists a positive constant L such that

$$||Ax - Ay|| \leq L||x - y||$$

for all $x, y \in U$. Thus, contraction operators are Lipschitz continuous operators with Lipschitz constant less than one.

Definition 3.43 *An element x of a normed space X is called a* fixed point *of an operator $A : U \subset X \to X$ if*

$$Ax = x.$$

Theorem 3.44 *Each contraction operator has at most one fixed point.*

Proof. Assume that x and y are two different fixed points of the contraction operator A. Then

$$0 \neq ||x - y|| = ||Ax - Ay|| \leq q||x - y||,$$

whence $1 \leq q$ follows. This is a contradiction to the fact that A is a contraction operator. □

Theorem 3.45 (Banach) *Let U be a complete subset of a normed space X and let $A : U \to U$ be a contraction operator. Then A has a unique fixed point.*

Proof. Starting from an arbitrary element $x_0 \in U$ we define a sequence (x_n) in U by the recursion

$$x_{n+1} := Ax_n, \quad n = 0, 1, 2, \ldots.$$

Then we have

$$\|x_{n+1} - x_n\| = \|Ax_n - Ax_{n-1}\| \le q\|x_n - x_{n-1}\|,$$

and from this we deduce by induction that

$$\|x_{n+1} - x_n\| \le q^n\|x_1 - x_0\|, \quad n = 1, 2, \ldots.$$

Hence, for $m > n$, by the triangle inequality and the geometric series it follows that

$$\begin{aligned} \|x_n - x_m\| &\le \|x_n - x_{n+1}\| + \|x_{n+1} - x_{n+2}\| + \cdots + \|x_{m-1} - x_m\| \\ &\le (q^n + q^{n+1} + \cdots + q^{m-1})\|x_1 - x_0\| \le \frac{q^n}{1-q}\,\|x_1 - x_0\|. \end{aligned} \tag{3.12}$$

Since $q^n \to 0$, $n \to \infty$, this implies that (x_n) is a Cauchy sequence, and therefore because U is complete there exists an element $x \in U$ such that $x_n \to x$, $n \to \infty$. Finally, the continuity of A from Remark 3.42 yields

$$x = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} Ax_n = Ax;$$

i.e., x is a fixed point of A. That this fixed point is unique we have already settled by Theorem 3.44. □

The main importance of Banach's fixed point theorem in numerical analysis originates from its constructive proof. Besides establishing existence of a fixed point by the method of successive approximations, it also provides an algorithm for obtaining numerical approximations. And this algorithm is very easy to program because of its iterative nature. We explicitly state this in the following theorem.

Theorem 3.46 *Let A be a contraction operator with contraction constant q mapping a complete subset U of a normed space X into itself. Then the* successive approximations

$$x_{n+1} := Ax_n, \quad n = 0, 1, 2, \ldots,$$

with arbitrary $x_0 \in U$ converge to the unique fixed point x of A. We have the a priori *error estimate*

$$\|x_n - x\| \le \frac{q^n}{1-q}\,\|x_1 - x_0\|$$

and the a posteriori *error estimate*

$$\|x_n - x\| \le \frac{q}{1-q}\,\|x_n - x_{n-1}\|$$

for all $n \in \mathbb{N}$.

Proof. The a priori error estimate follows from (3.12) by passing to the limit $m \to \infty$. The a posteriori estimate follows from the a priori estimate applied with starting element $x_0 = x_{n-1}$. □

The a priori estimate is used in order to obtain upper bounds on the number of iteration steps, which are necessary to achieve a desired accuracy. In order to guarantee that

$$\|x_n - x\| \le \varepsilon$$

for a given accuracy ε, by the a priori estimate we need

$$n \ge \frac{\ln \tilde{\varepsilon}}{\ln q}$$

iterations, where $\tilde{\varepsilon} = (1-q)\varepsilon/\|x_1 - x_0\|$. The smaller the contraction constant q, the fewer iteration steps are required. The a posteriori estimate, which in general yields better estimates as compared with the a priori estimate, is used to check the accuracy during the computation and terminate the iterations when the required accuracy is reached.

The property

$$\|Ax - Ay\| < \|x - y\|$$

for all x, y with $x \neq y$, which is weaker than the contraction property, is not sufficient in general to ensure the existence of a fixed point, as illustrated in the following example (see also Problem 3.18).

Example 3.47 The function $f : [0,\infty) \to [0,\infty)$ given by

$$f(x) := x + \frac{1}{1+x}$$

as a consequence of

$$f(x) - f(y) = \frac{x + y + xy}{1 + x + y + xy}\,(x - y)$$

fulfills the condition

$$|f(x) - f(y)| < |x - y|$$

for $x \neq y$. However, because of

$$\frac{1}{1+x} > 0$$

for all $x \ge 0$, it does not have a fixed point. □

We conclude this section by considering the special case of linear operators, i.e., by considering the *Neumann series* (see Problem 3.16).

Let $A : X \to Y$ be an operator mapping a set X into a set Y. If for each $y \in A(X)$ there is only one element $x \in X$ with $Ax = y$, then A is said to be *injective* and to have an inverse $A^{-1} : A(X) \to X$ defined by $A^{-1}y := x$. The inverse mapping satisfies $A^{-1}A = I$ on X and $AA^{-1} = I$ on $A(X)$, where I denotes the identity operator mapping each element into itself. If $A(X) = Y$, then the mapping is said to be *surjective.* The mapping is called *bijective* if it is injective and surjective, i.e., if the inverse mapping $A^{-1} : Y \to X$ exists.

Theorem 3.48 *Let $B : X \to X$ be a bounded linear operator on a Banach space X with $\|B\| < 1$, and let $I : X \to X$ denote the identity operator. Then $I - B$ is bijective; i.e., for each $z \in X$ the equation*

$$x - Bx = z$$

has a unique solution $x \in X$. The successive approximations

$$x_{n+1} := Bx_n + z, \quad n = 0, 1, 2, \dots,$$

with arbitrary $x_0 \in X$ converge to this solution, and we have the a priori *estimate*

$$\|x_n - x\| \le \frac{\|B\|^n}{1 - \|B\|} \|x_1 - x_0\|$$

and the a posteriori *estimate*

$$\|x_n - x\| \le \frac{\|B\|}{1 - \|B\|} \|x_n - x_{n-1}\|$$

for all $n \in \mathbb{N}$. Furthermore, the inverse operator $(I - B)^{-1}$ is bounded by

$$\|(I - B)^{-1}\| \le \frac{1}{1 - \|B\|}\,.$$

Proof. For fixed, but arbitrary, $z \in X$ we define the operator $A : X \to X$ by

$$Ax := Bx + z, \quad x \in X.$$

Then we have

$$\|Ax - Ay\| = \|B(x - y)\| \le \|B\|\,\|x - y\|$$

for all $x, y \in X$; i.e., A is a contraction with contraction number $q = \|B\|$. Now the statements of the theorem can be deduced from Theorem 3.46.

With the starting element $x_0 = z$ the successive approximations lead to

$$x_n = \sum_{k=0}^{n} B^k z$$

with the iterated operators $B^k : X \to X$ defined recursively by $B^0 := I$ and $B^k := BB^{k-1}$ for $k \in \mathbb{N}$. Hence, in view of Remark 3.25, we have

$$||x_n|| \le \sum_{k=0}^{n} ||B^k z|| \le \sum_{k=0}^{n} ||B||^k ||z|| \le \frac{||z||}{1 - ||B||},$$

and therefore, since $x_n \to (I - B)^{-1} z$, $n \to \infty$, it follows that

$$||(I - B)^{-1} z|| \le \frac{||z||}{1 - ||B||}$$

for all $z \in X$. □

3.7 Best Approximation

Definition 3.49 *Let $U \subset X$ be a subset of a normed space X and let $w \in X$. An element $v \in U$ is called a* best approximation *to w with respect to U if*

$$||w - v|| = \inf_{u \in U} ||w - u||,$$

i.e., if $v \in U$ has smallest distance from w.

Theorem 3.50 *Let U be a finite-dimensional subspace of a normed space X. Then for every element in X there exists a best approximation with respect to U.*

Proof. Let $w \in X$ and choose a minimizing sequence (u_n) for w; i.e., $u_n \in U$ satisfies

$$||w - u_n|| \to d := \inf_{u \in U} ||w - u||, \quad n \to \infty.$$

Because of $||u_n|| \le ||w - u_n|| + ||w||$ the sequence (u_n) is bounded. By Theorem 3.11 the sequence (u_n) contains a convergent subsequence $(u_{n(\ell)})$ with limit $v \in U$. Then

$$||w - v|| = \lim_{\ell \to \infty} ||w - u_{n(\ell)}|| = d$$

completes the proof. □

Theorem 3.51 *Let U be a linear subspace of a pre-Hilbert space X. An element v is a best approximation to $w \in X$ with respect to U if and only if*

$$(w - v, u) = 0 \tag{3.13}$$

for all $u \in U$, i.e., if and only if $w - v \perp U$. To each $w \in X$ there exists at most one best approximation with respect to U.

Proof. We begin by noting the equality

$$\|w-u\|^2 = \|w-v\|^2 + 2\,\mathrm{Re}(w-v, v-u) + \|v-u\|^2, \tag{3.14}$$

which is valid for all $u, v \in U$. From this, sufficiency of the condition (3.13) is obvious, since U is a linear subspace.

To establish the necessity we assume that v is a best approximation and $(w - v, u_0) \neq 0$ for some $u_0 \in U$. Then, since U is a linear subspace, we may assume that $(w - v, u_0) \in \mathbb{R}$. Choosing

$$u = v + \frac{(w-v, u_0)}{\|u_0\|^2}\, u_0,$$

from (3.14) we arrive at

$$\|w-u\|^2 = \|w-v\|^2 - \frac{(w-v,u_0)^2}{\|u_0\|^2} < \|w-v\|^2,$$

which contradicts the fact that v is a best approximation of w.

Finally, assume that v_1 and v_2 are best approximations. Then from (3.13) it follows that $(w - v_1, v_1 - v_2) = 0 = (w - v_2, v_1 - v_2)$. This implies $(v_1 - v_2, v_1 - v_2) = 0$, whence $v_1 = v_2$ follows. □

Theorem 3.52 *Let U be a complete linear subspace of a pre-Hilbert space X. Then to each element $w \in X$ there exists a unique best approximation with respect to U. The operator $P : X \to U$ mapping $w \in X$ onto its best approximation is a bounded linear operator with the properties*

$$P^2 = P \quad \textit{and} \quad \|P\| = 1.$$

It is called the orthogonal projection *from X onto U.*

Proof. Choose a sequence (u_n) with

$$\|w-u_n\|^2 \le d^2 + \frac{1}{n}\,, \quad n \in \mathbb{N}, \tag{3.15}$$

where $d := \inf_{u \in U} \|w - u\|$. Then

$$\|(w-u_n)+(w-u_m)\|^2 + \|u_n - u_m\|^2 = 2\|w-u_n\|^2 + 2\|w-u_m\|^2$$

$$\le 4d^2 + \frac{2}{n} + \frac{2}{m}$$

for all $n, m \in \mathbb{N}$, and since $\frac{1}{2}\,(u_n + u_m) \in U$, it follows that

$$\|u_n - u_m\|^2 \le 4d^2 + \frac{2}{n} + \frac{2}{m} - 4\left\|w - \frac{1}{2}\,(u_n+u_m)\right\|^2 \le \frac{2}{n} + \frac{2}{m}\,.$$

Hence, (u_n) is a Cauchy sequence, and since U is complete, there exists an element $v \in U$ such that $u_n \to v$, $n \to \infty$. Passing to the limit $n \to \infty$ in (3.15) shows that v is a best approximation of w with respect to U. Uniqueness of the best approximation follows from Theorem 3.51.

Trivially, we have $Pu = u$ for all $u \in U$, and this implies $P^2 = P$. From (3.13) it can be deduced that P is a linear operator and that

$$||w||^2 = ||Pw||^2 + ||w - Pw||^2 \geq ||Pw||^2$$

for all $w \in X$. Therefore, P is bounded with $||P|| \leq 1$. From Remark 3.25 and $P^2 = P$ it follows that $||P|| \geq 1$, which concludes the proof. □

Corollary 3.53 *Let U be a finite-dimensional linear subspace of a pre-Hilbert space X with basis $u_1, \dots, u_n$. The linear combination*

$$v = \sum_{k=1}^{n} \alpha_k u_k$$

is the best approximation to $w \in X$ with respect to U if and only if the coefficients $\alpha_1, \dots, \alpha_n$ satisfy the normal equations

$$\sum_{k=1}^{n} \alpha_k (u_k, u_j) = (w, u_j), \quad j = 1, \dots, n. \tag{3.16}$$

Proof. The normal equations (3.16) obviously are equivalent to (3.13). □

The normal equations for the best approximation in pre-Hilbert spaces provide further examples of systems of linear equations. The solution becomes trivial if the basis $u_1, \dots, u_n$ is orthonormal.

Corollary 3.54 *Let U be a finite-dimensional linear subspace of a pre-Hilbert space X with orthonormal basis $u_1, \dots, u_n$. Then the orthogonal projection operator is given by*

$$Pw = \sum_{k=1}^{n} (w, u_k) u_k, \quad w \in X.$$

Proof. This is trivial from either the orthogonality condition of Theorem 3.51 or the normal equations of Corollary 3.53. □

Problems

3.1 Show that (3.1) defines a norm on $\mathbb{C}^n$ for $p \geq 1$ and that

$$\lim_{p \to \infty} ||x||_p = ||x||_\infty$$

for all $x \in \mathbb{C}^n$.

3.2 Indicate the closed balls $\{x \in \mathbb{R}^2 : \|x\|_p \leq 1\}$ for $p = 1, 2, \infty$. What properties do they have in common?

3.3 Show that (3.1) does not define a norm on $\mathbb{C}^n$ for $0 < p < 1$.

3.4 For the ℓ_1 and ℓ_∞ norms on $\mathbb{C}^n$ show that $\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty$.

3.5 Let X and Y be normed spaces with norms $\|\cdot\|_X$ and $\|\cdot\|_Y$, respectively. Show that

$$\|(x, y)\| := \|x\|_X + \|y\|_Y,$$

$$\|(x, y)\| := (\|x\|_X^2 + \|y\|_Y^2)^{1/2},$$

$$\|(x, y)\| := \max(\|x\|_X, \|y\|_Y),$$

for $(x, y) \in X \times Y$ define norms on the product $X \times Y$.

3.6 Show that convergent sequences are bounded.

3.7 Let (x_n) be a sequence of elements of a normed space X. The *series*

$$\sum_{k=1}^{\infty} x_k$$

is called *convergent* if the sequence (S_n) of partial sums

$$S_n := \sum_{k=1}^{n} x_k$$

converges. The limit $S = \lim_{n\to\infty} S_n$ is called the *sum* of the series. Show that in a Banach space X the convergence of the series

$$\sum_{k=1}^{\infty} \|x_k\|$$

is a sufficient condition for the convergence of the series $\sum_{k=1}^{\infty} x_k$ and that

$$\left\| \sum_{k=1}^{\infty} x_k \right\| \leq \sum_{k=1}^{\infty} \|x_k\|.$$

3.8 A norm $\|\cdot\|_a$ on a linear space X is called *stronger* than a norm $\|\cdot\|_b$ if every sequence converging with respect to the norm $\|\cdot\|_a$ also converges with respect to the norm $\|\cdot\|_b$. Show that $\|\cdot\|_a$ is stronger than $\|\cdot\|_b$ if and only if there exists a positive number C such that $\|x\|_b \leq C\|x\|_a$ for all $x \in X$. Show that on $C[a, b]$ the maximum norm is stronger than the L_2 norm (and stronger than the L_1 norm). Construct a counterexample to demonstrate that the maximum norm and the L_2 norm (and the maximum norm and the L_1 norm) are not equivalent.

3.9 Show that in a normed space the operations of addition and multiplication by a scalar are continuous functions. Show that in a pre-Hilbert space the scalar product is a continuous function.

3.10 Show that a norm $\|.\|$ on a linear space X is generated by a scalar product if and only if the parallelogram equality

$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2)$$

holds for all $x, y \in X$. Show that the ℓ_1 and ℓ_∞ norms on $\mathbb{C}^n$ are not generated by scalar products.

3.11 Let A be a positive definite $n \times n$ matrix and denote by $(\cdot\,,\cdot)$ the Euclidean scalar product on $\mathbb{C}^n$. Show that (Ax, y) defines a scalar product on $\mathbb{C}^n$.

3.12 Show that eigenvectors of a matrix for different eigenvalues are linearly independent.

3.13 Let X and Y be normed spaces and denote by $L(X, Y)$ the linear space of all bounded linear operators $A : X \to Y$. Show that $L(X, Y)$ equipped with

$$\|A\| := \sup_{\|x\|=1} \|Ax\|$$

again is a normed space and that $L(X, Y)$ is a Banach space if Y is a Banach space.

3.14 Let $A : X \to X$ denote an operator from a normed space X into itself. The iterated operators $A^n : X \to X$, $n = 0, 1, \dots$, are defined recursively by $A^0 = I$ and $A^n := AA^{n-1}$ for $n \in \mathbb{N}$. If A is bounded and linear, show that $\|A^n\| \leq \|A\|^n$.

3.15 Show that for $n \times n$ matrices A the series

$$\sum_{k=0}^{\infty} \frac{1}{k!} A^k$$

converges (with respect to any norm on $\mathbb{C}^n$), and denote the sum of the series by e^A. Show that if λ is an eigenvalue of A, then e^λ is an eigenvalue of e^A.

3.16 Show that if $B : X \to X$ is a linear operator on a Banach space X with $\|B\| < 1$, then the *Neumann series*

$$\sum_{k=0}^{\infty} B^k = (I - B)^{-1}$$

converges in the Banach space $L(X, X)$.

3.17 Let U be a complete subset of a normed space X and let $A : U \to U$ be a continuous operator, and assume that A^m is a contraction for some $m \in \mathbb{N}$. Show that A has a unique fixed point and that the successive approximations $x_{n+1} := Ax_n$, $n = 0, 1, \dots$, with arbitrary $x_0 \in U$ converge to this fixed point.

3.18 A subset U of a normed space X is called *sequentially compact* if each sequence from U contains a convergent subsequence with limit in U. Let U be a

complete and sequentially compact subset of a normed space X and let $A : U \to U$ be an operator with the property

$$\|Ax - Ay\| < \|x - y\|$$

for all $x, y \in U$ with $x \neq y$. Show that A has a unique fixed point and that the successive approximations $x_{n+1} := Ax_n$, $n = 0, 1, \dots$, with arbitrary $x_0 \in U$ converge to this fixed point.

3.19 Let $\{u_n : n \in \mathbb{N}\}$ be an orthonormal system in a pre-Hilbert space X. Show that the following properties are equivalent:

(a) $\operatorname{span}\{u_n : n \in \mathbb{N}\}$ is dense in X

(b) Each $\varphi \in X$ can be expanded in a *Fourier series*

$$\varphi = \sum_{n=1}^{\infty} (\varphi, u_n) u_n$$

(c) For each $\varphi \in X$ we have *Parseval's equality*

$$\|\varphi\|^2 = \sum_{n=1}^{\infty} |(\varphi, u_n)|^2.$$

Show that properties (a)–(c) imply that

(d) $x = 0$ is the only element in X with $(x, u_n) = 0$ for all $n \in \mathbb{N}$,

and that (a), (b), (c), and (d) are equivalent if X is a Hilbert space.

3.20 Show that the best approximation to a function $f \in C[0, 2\pi]$ in the L^2 norm with respect to the space of trigonometric polynomials of degree at most n is given by the partial sum

$$(P_n f)(x) = \frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos kx + \sum_{k=1}^{n} b_k \sin kx, \quad x \in [0, 2\pi],$$

of the *Fourier series* of f with the Fourier coefficients

$$a_k = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos kx \, dx, \quad b_k = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin kx \, dx.$$

4

Iterative Methods for Linear Systems

This chapter is devoted to applying the analysis developed in the previous chapter to the iterative solution of systems of linear equations. In particular, we will discuss in detail the Jacobi and the Gauss–Seidel iterations, which essentially go back to Gauss. In *Supplementum Theoriae Combinationis Observationum Erroribus Minime Obnoxia*, published in 1822, Gauss used a variant of the Gauss–Seidel method for the solution of the linear systems arising through his least squares method, since they were too large for elimination methods.

With the advent of computers the size of the linear systems that could be solved grew enormously, leading to the requirement of speedup of the convergence of the classical Jacobi and Gauss–Seidel iterations. In this context, we will introduce the reader to the idea of relaxation methods, including a typical example that illustrates the dramatic gain in the speed of convergence by overrelaxation. We will conclude the section with the idea of defect correction iteration and indicate its application to the very efficient solution of the large linear systems arising from the discretization of linear differential and integral equations by two-grid and multigrid methods.

4.1 Jacobi and Gauss–Seidel Iterations

We start by supplementing the sufficient condition of Theorem 3.48 for convergence of the method of successive approximations by establishing a necessary and sufficient condition for the finite-dimensional case.

Theorem 4.1 *Let B be an $n \times n$ matrix. Then the successive approximations*

$$x_{\nu+1} := Bx_\nu + z, \quad \nu = 0, 1, 2, \dots,$$

converge for each $z \in \mathbb{C}^n$ and each $x_0 \in \mathbb{C}^n$ if and only if

$$\rho(B) < 1$$

for the spectral radius of B.

Proof. If $\rho(B) < 1$, then by Theorem 3.32 there exists a norm $\|\cdot\|$ on $\mathbb{C}^n$ such that $\|B\| < 1$. Now convergence follows from Theorem 3.48 together with the equivalence of all norms on $\mathbb{C}^n$ according to Theorem 3.8.

Conversely, suppose that convergence holds. If we assume that $\rho(B) \geq 1$, then there exists an eigenvalue λ of B with $|\lambda| \geq 1$. Let x denote an associated eigenvector. Then the successive iterations for the right-hand side $z = x$ and the starting element $x_0 = x$ lead to the divergent sequence $x_\nu = \left(\sum_{k=0}^{\nu} \lambda^k\right) x$. This is a contradiction. □

We note that Theorem 4.1 remains valid for bounded linear operators $B : X \to X$ in infinite-dimensional Banach spaces with the definition of the spectral radius appropriately modified. However, the proof requires a different and deeper analysis.

For the iterative solution of a system of linear equations of the form

$$Ax = y$$

we distinguish different methods by the way in which the original system is transformed into an equivalent fixed-point form. We decompose A by

$$A = D + A_L + A_R$$

into a diagonal matrix

$$D = \operatorname{diag}(a_{11}, \dots, a_{nn}),$$

a proper lower (left) triangular matrix

$$A_L = \begin{pmatrix} 0 & & & & & \\ a_{21} & 0 & & & & \\ a_{31} & a_{32} & 0 & & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & \cdot & \cdot & a_{n,n-1} & 0 \end{pmatrix},$$

and a proper upper (right) triangular matrix

$$A_R = \begin{pmatrix} 0 & a_{12} & a_{13} & \cdot & \cdot & a_{1n} \\ & 0 & a_{23} & \cdot & \cdot & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & 0 & a_{n-1,n} \\ & & & & & 0 \end{pmatrix}.$$

We assume that all the diagonal entries of A are different from zero. Hence the inverse D^{-1} of D exists.

In the method attributed to Jacobi, which is sometimes also called the method of simultaneous displacements, the system $Ax = y$ is transformed into the equivalent form

$$x = -D^{-1}(A_L + A_R)x + D^{-1}y,$$

and the latter is solved by successive approximations

$$x_{\nu+1} := -D^{-1}(A_L + A_R)x_\nu + D^{-1}y, \quad \nu = 0, 1, 2, \ldots,$$

with arbitrarily chosen starting element x_0. Written in components, one step of the Jacobi iteration scheme reads

$$x_{\nu+1,j} = -\sum_{\substack{k=1 \\ k \neq j}}^{n} \frac{a_{jk}}{a_{jj}} x_{\nu,k} + \frac{y_j}{a_{jj}}, \quad j = 1, \ldots, n.$$

Theorem 4.2 *Assume that the matrix $A = (a_{jk})$ satisfies*

$$q_\infty := \max_{j=1,\ldots,n} \sum_{\substack{k=1 \\ k \neq j}}^{n} \left| \frac{a_{jk}}{a_{jj}} \right| < 1 \tag{4.1}$$

or

$$q_1 := \max_{k=1,\ldots,n} \sum_{\substack{j=1 \\ j \neq k}}^{n} \left| \frac{a_{jk}}{a_{jj}} \right| < 1 \tag{4.2}$$

or

$$q_2 := \left(\sum_{\substack{j,k=1 \\ j \neq k}}^{n} \left| \frac{a_{jk}}{a_{jj}} \right|^2 \right)^{1/2} < 1. \tag{4.3}$$

Then the Jacobi method, *or* method of simultaneous displacements,

$$x_{\nu+1,j} = -\sum_{\substack{k=1 \\ k \neq j}}^{n} \frac{a_{jk}}{a_{jj}} x_{\nu,k} + \frac{y_j}{a_{jj}}, \quad j = 1, \ldots, n, \quad \nu = 0, 1, 2, \ldots,$$

converges for each $y \in \mathbb{C}^n$ and each $x_0 \in \mathbb{C}^n$ to the the unique solution of $Ax = y$ (in any norm on $\mathbb{C}^n$). For $\mu = 1, 2, \infty$, if $q_\mu < 1$, we have the a priori error estimate

$$\|x_\nu - x\|_\mu \leq \frac{q_\mu^\nu}{1 - q_\mu} \|x_1 - x_0\|_\mu$$

and the a posteriori error estimate

$$\|x_\nu - x\|_\mu \leq \frac{q_\mu}{1 - q_\mu} \|x_\nu - x_{\nu-1}\|_\mu$$

for all $\nu \in \mathbb{N}$.

Proof. The Jacobi matrix $-D^{-1}(A_L + A_R)$ has diagonal entries zero and off-diagonal entries $-a_{jk}/a_{jj}$. Hence by Theorem 3.26 we have

$$\| - D^{-1}(A_L + A_R)\|_\infty = q_\infty,$$

$$\| - D^{-1}(A_L + A_R)\|_1 = q_1,$$

$$\| - D^{-1}(A_L + A_R)\|_2 \leq q_2.$$

Now the assertion follows from Theorem 3.48. □

Note that the sufficient convergence conditions (4.1)–(4.3) are not equivalent. Roughly speaking, each criterion ensures convergence if the diagonal entries of A are dominant. The condition (4.1) can also be written as

$$\sum_{\substack{k=1 \\ k \neq j}}^{n} |a_{jk}| < |a_{jj}|, \quad j = 1, \ldots, n; \tag{4.4}$$

i.e., the matrix A is required to be *strictly row-diagonally dominant.* From (4.2) it can be deduced (see Problem 4.4) that if

$$\sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{jk}| < |a_{kk}|, \quad k = 1, \ldots, n, \tag{4.5}$$

i.e., if the matrix A is *strictly column-diagonally dominant*, then the Jacobi iterations converge.

For the Gauss–Seidel method, which is also known as the method of successive displacements, we proceed differently and transform $Ax = y$ via

$$(D + A_L)x = -A_R x + y$$

into the equivalent form

$$x = -(D + A_L)^{-1} A_R x + (D + A_L)^{-1} y,$$

which is then solved by the successive approximations

$$x_{\nu+1} := -(D + A_L)^{-1} A_R x_\nu + (D + A_L)^{-1} y, \quad \nu = 0, 1, 2, \ldots,$$

with arbitrarily chosen starting element x_0. For the actual computations we rewrite this as

$$(D + A_L)x_{\nu+1} = -A_R x_\nu + y, \quad \nu = 0, 1, 2, \ldots,$$

and solve the linear system for $x_{\nu+1}$ with the lower triangular matrix $D+A_L$ by forward substitution. This leads to the Gauss–Seidel iteration scheme in the following explicit form:

$$x_{\nu+1,j} = -\sum_{k=1}^{j-1} \frac{a_{jk}}{a_{jj}} x_{\nu+1,k} - \sum_{k=j+1}^{n} \frac{a_{jk}}{a_{jj}} x_{\nu,k} + \frac{y_j}{a_{jj}}, \quad j = 1, \ldots, n.$$

Here and in the sequel empty sums have to be interpreted as zero.

In the Jacobi iteration scheme all the components of the new approximation vector $x_{\nu+1}$ are obtained by using only the components of the previous approximation vector x_ν, which explains why this method is also called the method of simultaneous displacements. However, in the Gauss–Seidel iterations each new component of $x_{\nu+1}$ is immediately used in the computation of the next component; i.e., for computing the jth component $x_{\nu+1,j}$, the values $x_{\nu+1,1}, x_{\nu+1,2}, \ldots, x_{\nu+1,j-1}$ are already used. This is very convenient for computer calculations, since the new values can be stored in the locations held by the old values, which reduces the storage requirements.

Theorem 4.3 *Assume that the matrix* $A = (a_{jk})$ *fulfills the* Sassenfeld criterion

$$p := \max_{j=1,\ldots,n} p_j < 1,$$

where the numbers p_j *are recursively defined by*

$$p_1 := \sum_{k=2}^{n} \left| \frac{a_{1k}}{a_{11}} \right|, \quad p_j := \sum_{k=1}^{j-1} \left| \frac{a_{jk}}{a_{jj}} \right| p_k + \sum_{k=j+1}^{n} \left| \frac{a_{jk}}{a_{jj}} \right|, \quad j = 2, \ldots, n.$$

Then the Gauss–Seidel method, *or* method of successive displacements,

$$x_{\nu+1,j} = -\sum_{k=1}^{j-1} \frac{a_{jk}}{a_{jj}} x_{\nu+1,k} - \sum_{k=j+1}^{n} \frac{a_{jk}}{a_{jj}} x_{\nu,k} + \frac{y_j}{a_{jj}}, \; j = 1, \ldots, n, \; \nu = 0, 1, 2, \ldots,$$

converges for each $y \in \mathbb{C}^n$ *and each* $x_0 \in \mathbb{C}^n$ *to the the unique solution of* $Ax = y$ *(in any norm on* $\mathbb{C}^n$*). We have the a priori error estimate*

$$\|x_\nu - x\|_\infty \le \frac{p^\nu}{1-p} \|x_1 - x_0\|_\infty$$

and the a posteriori error estimate

$$\|x_\nu - x\|_\infty \le \frac{p}{1-p} \|x_\nu - x_{\nu-1}\|_\infty$$

for all $\nu \in \mathbb{N}$.

Proof. Consider the equation

$$(D + A_L)x = -A_R z$$

for $z \in \mathbb{C}^n$ with $\|z\|_\infty = 1$, that is,

$$x_j = -\sum_{k=1}^{j-1} \frac{a_{jk}}{a_{jj}} x_k - \sum_{k=j+1}^{n} \frac{a_{jk}}{a_{jj}} z_k, \quad j = 1, \dots, n.$$

By induction, this implies that $|x_j| \le p_j$ for $j = 1, \dots, n$, and therefore $\|x\|_\infty \le p$. Hence we have

$$\|(D + A_L)^{-1} A_R\|_\infty \le p,$$

and the assertion of the theorem follows from Theorem 3.48. □

Corollary 4.4 *Assume that the matrix A is strictly row-diagonally dominant. Then the Gauss–Seidel iterations converge.*

Example 4.5 *The tridiagonal matrix*

$$A = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}$$

from Example 2.1 is not strictly row-diagonally dominant, but it satisfies the Sassenfeld criterion.

Proof. Obviously, $q_\infty = 1$; i.e., (4.1) is not fulfilled. We have the recursion

$$p_1 = \frac{1}{2}, \quad p_j = \frac{1}{2} p_{j-1} + \frac{1}{2}, \quad j = 2, \dots, n-1, \quad p_n = \frac{1}{2} p_{n-1}.$$

From this, by induction, it follows that

$$p_j = 1 - \frac{1}{2^j}, \quad j = 1, \dots, n-1, \quad p_n = \frac{1}{2} - \frac{1}{2^n}.$$

Therefore,

$$p = 1 - \frac{1}{2^{n-1}} < 1,$$

and this implies convergence of the Gauss–Seidel iterations by Theorem 4.3. □

Since the matrix A is tridiagonal, the system $Ax = y$ can be solved efficiently by elimination (see Problem 2.9). Nevertheless, this matrix provides a very suitable example for the analysis of iterative methods for linear systems arising in the discretization of ordinary and partial differential equations. This is due to the fact that in more general cases, for example for the linear system of Example 2.2, there are more technical details to consider, which distract from the basic principles. However, these basic principles do not depend on the dimension of the underlying differential equation problem.

In Example 4.5, if n is large, the contraction number p will be close to one, i.e., the convergence rate of the Gauss–Seidel iterations will be unsatisfactorily slow. Before we indicate how the convergence can be accelerated, we continue by discussing a weaker form of row-diagonal dominance.

Definition 4.6 *An $n \times n$ matrix $A = (a_{jk})$ is called* reducible *if there exist two nonempty sets $N, M \subset \{1, \dots, n\}$ such that*

$$N \cap M = \emptyset, \quad N \cup M = \{1, \dots, n\},$$

and

$$a_{jk} = 0, \quad j \in N, \ k \in M.$$

Otherwise the matrix is called irreducible.

A reducible matrix A, after a reordering of the rows and columns, can be partitioned into a 2×2 block matrix of the form

$$A = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}$$

(see Problem 4.5). Therefore, solving a linear system with the matrix A can be reduced to solving two smaller linear systems with the matrices A_{11} and A_{22}.

Theorem 4.7 *Assume that the matrix $A = (a_{jk})$ is irreducible and* weakly row-diagonally dominant*; i.e., A is row-diagonally dominant,*

$$\sum_{\substack{k=1 \\ k \neq j}}^{n} |a_{jk}| \leq |a_{jj}|, \quad j = 1, \dots, n, \tag{4.6}$$

with inequality holding for at least one row j. Then the Jacobi iterations converge for each $y \in \mathbb{C}^n$ and each $x_0 \in \mathbb{C}^n$ to the unique solution of $Ax = y$ (in any norm on $\mathbb{C}^n$).

Proof. By (4.6) and Theorem 3.26 we have that $\|B\|_\infty \leq 1$ for the Jacobi matrix $B = -D^{-1}(A_L + A_R)$. Therefore, from Theorem 3.32 it follows that $\rho(B) \leq 1$ for the spectral radius.

Now assume that there exists an eigenvalue λ of B with $|\lambda| = 1$. For the associated eigenvector we may assume that $\|x\|_\infty = 1$. Then from $\lambda x = Bx$ we obtain the inequality

$$|\lambda|\,|x_j| \le \sum_{\substack{k=1\\k\neq j}}^{n} \left|\frac{a_{jk}}{a_{jj}}\right| |x_k| \le \sum_{\substack{k=1\\k\neq j}}^{n} \left|\frac{a_{jk}}{a_{jj}}\right| \le 1, \quad j = 1,\dots,n. \tag{4.7}$$

Let $N := \{j : |x_j| = 1\}$. Since $\|x\|_\infty = 1$, we have that $N \neq \emptyset$. For $j \in N$ we have $|\lambda|\,|x_j| = 1$, and therefore equality holds in (4.7); i.e.,

$$\sum_{\substack{k=1\\k\neq j}}^{n} \left|\frac{a_{jk}}{a_{jj}}\right| = 1, \quad j \in N.$$

From this it follows that

$$M := \{1,\dots,n\} \setminus N \neq \emptyset,$$

since A is weakly row-diagonally dominant. Because A is irreducible, there exists $j_0 \in N$ and $k_0 \in M$ such that $a_{j_0k_0} \neq 0$. Now by using

$$|a_{j_0k_0}|\,|x_{k_0}| < |a_{j_0k_0}|$$

we obtain the contradiction

$$1 = |x_{j_0}| = |\lambda|\,|x_{j_0}| = \sum_{\substack{k=1\\k\neq j}}^{n} \left|\frac{a_{j_0k}}{a_{j_0j_0}}\right| |x_k| < \sum_{\substack{k=1\\k\neq j}}^{n} \left|\frac{a_{j_0k}}{a_{j_0j_0}}\right| \le 1.$$

Therefore, we have $\rho(B) < 1$, and the statement of the theorem follows from Theorem 4.1. □

We leave it to the reader as an exercise to show that the matrix A from Example 4.5 is irreducible and weakly row-diagonally dominant (see Problem 4.6), implying convergence of the Jacobi iterations.

4.2 Relaxation Methods

From combining the a priori error estimate of Theorem 3.48 with Theorem 4.1 we see that the spectral radius $\rho(B)$ of the iteration matrix B may be considered as a measure for the speed of convergence of the successive approximations. Therefore, it is desirable to design the iterative scheme such that $\rho(B)$ becomes small. This aim is the motivation of the relaxation methods to be discussed in this section.

Each step of the Jacobi iterations can be written in the form

$$x_{\nu+1} = x_\nu + D^{-1}(y - Ax_\nu),$$

indicating how the new approximation $x_{\nu+1}$ is obtained by correcting the previous approximation x_ν. The basic idea of the relaxation methods is to multiply the correction term by some weight factor. Note that if the following relaxation iterations converge, then they converge to a solution of $Ax = y$.

Definition 4.8 *The iterative scheme*

$$x_{\nu+1} := x_\nu + \omega D^{-1}(y - Ax_\nu), \quad \nu = 0, 1, 2, \ldots,$$

i.e., in components

$$x_{\nu+1,j} = x_{\nu,j} + \frac{\omega}{a_{jj}}\left[y_j - \sum_{k=1}^{n} a_{jk}x_{\nu,k}\right], \quad j = 1, \ldots, n,$$

is known as the Jacobi method with relaxation. *The weight factor* $\omega > 0$ *is called the* relaxation parameter.

Theorem 4.9 *Assume that the Jacobi matrix* $B := -D^{-1}(A_L + A_R)$ *has real eigenvalues and spectral radius less than one. Then the spectral radius of the iteration matrix*

$$I - \omega D^{-1}A = (1 - \omega)I - \omega D^{-1}(A_L + A_R)$$

for the Jacobi method with relaxation becomes minimal for the relaxation parameter

$$\omega_{\text{opt}} = \frac{2}{2 - \lambda_{\max} - \lambda_{\min}}$$

and has spectral radius

$$\rho(I - \omega_{\text{opt}}D^{-1}A) = \frac{\lambda_{\max} - \lambda_{\min}}{2 - \lambda_{\max} - \lambda_{\min}},$$

where $\lambda_{\min}$ *and* $\lambda_{\max}$ *denote the smallest and the largest eigenvalue of* B, *respectively. In the case* $\lambda_{\min} \neq -\lambda_{\max}$ *the convergence of the Jacobi method with optimal relaxation parameter is faster than the convergence of the Jacobi method without relaxation.*

Proof. For $\omega > 0$ the equation $Bu = \lambda u$ is equivalent to

$$[(1 - \omega)I + \omega B]u = [1 - \omega + \omega\lambda]u.$$

Hence the eigenvalues λ of B correspond to the eigenvalues $1 - \omega + \omega\lambda$ of $(1 - \omega)I + \omega B$. Therefore, the eigenvalues of $(1 - \omega)I + \omega B$ are real, and

the smallest eigenvalue of $(1-\omega)I+\omega B$ is given by $1-\omega+\omega\lambda_{\min}$ and the largest by $1-\omega+\omega\lambda_{\max}$. Obviously, the spectral radius becomes minimal if the smallest and the largest eigenvalue are of opposite sign and have the same absolute value, i.e., if

$$1-\omega_{\text{opt}}+\omega_{\text{opt}}\lambda_{\min}=-1+\omega_{\text{opt}}-\omega_{\text{opt}}\lambda_{\max}.$$

From this, elementary algebra yields the optimal parameter ω_{opt} and the spectral radius $\rho(I-\omega_{\text{opt}}D^{-1}A)$ as stated in the theorem. □

For the Gauss–Seidel iterations, from $(D+A_L)x_{\nu+1}=-A_Rx_\nu+y$ it follows that

$$x_{\nu+1}=x_\nu+D^{-1}[y-A_Lx_{\nu+1}-(D+A_R)x_\nu].$$

Hence, the corresponding relaxation method is defined as follows. Note again that if the relaxation iterations converge, then they converge to a solution of $Ax=y$.

Definition 4.10 *The iterative scheme*

$$x_{\nu+1}=x_\nu+\omega D^{-1}[y-A_Lx_{\nu+1}-(D+A_R)x_\nu],\quad \nu=0,1,2,\dots,$$

i.e., in components

$$x_{\nu+1,j}=x_{\nu,j}+\frac{\omega}{a_{jj}}\left[y_j-\sum_{k=1}^{j-1}a_{jk}x_{\nu+1,k}-\sum_{k=j}^{n}a_{jk}x_{\nu,k}\right],\quad j=1,\dots,n,$$

is known as the Gauss–Seidel method with relaxation *or as the* successive overrelaxation (SOR) method *with relaxation coefficient* $\omega>0$.

From

$$(D+\omega A_L)x_{\nu+1}=\omega y+[(1-\omega)D-\omega A_R]x_\nu$$

we obtain that the iteration matrix of the SOR method is given by

$$B(\omega):=(D+\omega A_L)^{-1}[(1-\omega)D-\omega A_R].$$

Here, as opposed to the relaxation of the Jacobi method, the iteration matrix depends nonlinearly on the relaxation parameter. This makes the convergence analysis of the SOR method more complicated.

Theorem 4.11 (Kahan) *A necessary condition for the SOR method to be convergent is that* $0<\omega<2$.

Proof. Since the eigenvalues $\mu_1,\dots,\mu_n$ of $B(\omega)$ are the zeros of the characteristic polynomial, they satisfy

$$\prod_{j=1}^{n}\mu_j=\det B(\omega)$$

(where multiple eigenvalues are repeated according to their algebraic multiplicity). From this, by the multiplication rules for determinants and since $D + \omega A_L$ and $(1-\omega)D - \omega A_R$ are triangular matrices, it follows that

$$\prod_{j=1}^{n} \mu_j = \det(D + \omega A_L)^{-1} \det[(1-\omega)D - \omega A_R] = (1-\omega)^n.$$

This now implies

$$\rho[B(\omega)] \geq |1-\omega|,$$

and from Theorem 4.1 we conclude the necessity of $0 < \omega < 2$ for convergence. □

Theorem 4.12 (Ostrowski) *If A is Hermitian and positive definite, then the SOR method converges for all $x_0 \in \mathbb{C}^n$, all $y \in \mathbb{C}^n$, and all $0 < \omega < 2$ to the unique solution of $Ax = y$.*

Proof. Let μ be an eigenvalue of $B(\omega)$ with eigenvector x; i.e.,

$$[(1-\omega)D - \omega A_R]x = \mu(D + \omega A_L)x.$$

With the aid of

$$(2-\omega)D - \omega A - \omega(A_R - A_L) = 2[(1-\omega)D - \omega A_R]$$

and

$$(2-\omega)D + \omega A - \omega(A_R - A_L) = 2[D + \omega A_L]$$

we deduce that

$$[(2-\omega)D - \omega A - \omega(A_R - A_L)]x = \mu[(2-\omega)D + \omega A - \omega(A_R - A_L)]x.$$

Taking the Euclidean scalar product with x, it now follows that

$$\mu = \frac{(2-\omega)d - \omega a + i\omega s}{(2-\omega)d + \omega a + i\omega s},$$

where we have set

$$a := (Ax, x), \quad d := (Dx, x), \quad s := i(A_R x - A_L x, x).$$

Since A is positive definite, we have $a > 0$ and $d > 0$, and since A is Hermitean, s is real. From

$$|(2-\omega)d - \omega a| < |(2-\omega)d + \omega a|$$

for $0 < \omega < 2$ we now can conclude that $|\mu| < 1$ for $0 < \omega < 2$. Hence convergence of the SOR method for $0 < \omega < 2$ follows from Theorem 4.1. □

The calculation of the optimal relaxation parameter, i.e., the parameter minimizing the spectral radius, is difficult except in some simple cases. Usually it is obtained only approximately by trial and error, based on trying several values of ω and observing the effect on the speed of convergence. However, the effort is well worth the time, since the resulting improvement of the convergence can be considerably large, as we will indicate by the following analysis, which relates the convergence of the SOR method to that of the Jacobi method for a certain class of matrices that occurs in the discretization of boundary value problems.

Definition 4.13 *A matrix $A = D + A_L + A_R$ with nonsingular diagonal D is called* consistently ordered *if the eigenvalues of*

$$C(\alpha) := -\alpha D^{-1} A_L - \frac{1}{\alpha} D^{-1} A_R, \quad \alpha \in \mathbb{C} \setminus \{0\},$$

do not depend on α.

The following theorem ensures that the analysis we are going to develop applies to the matrix of Example 2.1, i.e., of Example 4.5.

Remark 4.14 *Tridiagonal matrices with nonzero diagonal elements are consistently ordered.*

Proof. After introducing the diagonal matrix

$$S(\alpha) := \operatorname{diag}(1, \alpha, \alpha^2, \ldots, \alpha^{n-1})$$

for tridiagonal matrices $A = D + A_L + A_R$, we have that

$$S(\alpha) C(1) S(\alpha)^{-1} = C(\alpha);$$

i.e., all matrices $C(\alpha)$ are similar, and therefore they have the same eigenvalues. □

Without going into detail, we wish to say that a much wider class of matrices arising in the discretization of differential equations enjoys the property of being consistently ordered in the sense of Definition 4.13. For a more comprehensive study we refer to [61, 63, 66].

Theorem 4.15 (Young) *Assume that A is a consistently ordered matrix and that the eigenvalues of the Jacobi matrix $-D^{-1}(A_L + A_R)$ are real with spectral radius $\Lambda = \rho[-D^{-1}(A_L + A_R)] < 1$. Then the SOR method converges for all $0 < \omega < 2$. The spectral radius of the SOR matrix $B(\omega)$ is minimal for*

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \Lambda^2}} \geq 1.$$

In this case we have

$$\rho[B(\omega_{\text{opt}})] = \frac{1 - \sqrt{1 - \Lambda^2}}{1 + \sqrt{1 - \Lambda^2}} \,.$$

Proof. From

$$(I + \omega D^{-1}A_L)[\mu I - B(\omega)] = \mu(I + \omega D^{-1}A_L) - D^{-1}[(1-\omega)D - \omega A_R]$$

$$= (\mu + \omega - 1)I + \sqrt{\mu}\,\omega\left(\sqrt{\mu}\,D^{-1}A_L + \frac{1}{\sqrt{\mu}}\,D^{-1}A_R\right)$$

and the fact that $I + \omega D^{-1}A_L$ is nonsingular it can be seen that $\mu \neq 0$ is an eigenvalue of $B(\omega)$ if and only if

$$\lambda = \frac{\mu + \omega - 1}{\sqrt{\mu}\,\omega} \tag{4.8}$$

is an eigenvalue of

$$-\sqrt{\mu}\,D^{-1}A_L - \frac{1}{\sqrt{\mu}}\,D^{-1}A_R.$$

Since A is assumed to be consistently ordered, it follows that $\mu \neq 0$ is an eigenvalue of $B(\omega)$ if and only if λ is an eigenvalue of $-D^{-1}(A_L + A_R)$.

Solving the quadratic equation

$$\mu + \omega - 1 = \sqrt{\mu}\,\omega\lambda$$

yields

$$\mu = \left(\frac{\omega\lambda}{2} \pm \sqrt{\frac{\omega^2\lambda^2}{4} + 1 - \omega}\right)^2.$$

Setting $\alpha = -1$ in Definition 4.13, it is obvious that if λ is an eigenvalue of $-D^{-1}(A_L + A_R)$, then $-\lambda$ also is an eigenvalue of $-D^{-1}(A_L + A_R)$. Therefore, since we are interested only in the spectral radius of $B(\omega)$, we can confine our considerations to

$$\mu = \left(\frac{\omega|\lambda|}{2} + \sqrt{\frac{\omega^2\lambda^2}{4} + 1 - \omega}\right)^2.$$

Because of $|\lambda| < 1$, the quadratic equation

$$\omega^2\lambda^2 - 4\omega + 4 = 0$$

has two real solutions, and only one of them belongs to the interval $(0,2)$, namely

$$\omega_0(\lambda) = \frac{2}{1 + \sqrt{1 - \lambda^2}} \geq 1.$$

This implies that

$$\omega^2\lambda^2 - 4\omega + 4 \geq 0, \quad 0 < \omega \leq \omega_0(\lambda).$$

Therefore, we have

$$|\mu(\omega)| = \left(\frac{\omega|\lambda|}{2} + \sqrt{\frac{\omega^2\lambda^2}{4} + 1 - \omega}\right)^2, \quad 0 < \omega \le \omega_0(\lambda). \tag{4.9}$$

For $\omega_0(\lambda) < \omega < 2$ the eigenvalues are complex, with

$$|\mu(\omega)| = \omega - 1, \quad \omega_0(\lambda) < \omega < 2. \tag{4.10}$$

From the expressions (4.9) and (4.10) it can be seen that $|\mu(\omega)|$ is monotonically nondecreasing with respect to $|\lambda|$. Hence

$$\rho[B(\omega)] = \begin{cases} \left(\dfrac{\omega\Lambda}{2} + \sqrt{\dfrac{\omega^2\Lambda^2}{4} + 1 - \omega}\right)^2, & 0 < \omega \le \omega_0(\Lambda), \\ \omega - 1, & \omega_0(\Lambda) < \omega < 2. \end{cases} \tag{4.11}$$

The function

$$f(\omega) := \frac{\omega\Lambda}{2} + \sqrt{\frac{\omega^2\Lambda^2}{4} + 1 - \omega}$$

has the properties $f(0) = 1$ and

$$f'(\omega) = \frac{\Lambda}{2} + \frac{\omega\Lambda^2 - 2}{2\sqrt{\omega^2\Lambda^2 + 4 - 4\omega}} < 0.$$

The latter follows from

$$\Lambda^2(4 - 4\omega + \omega^2\Lambda^2) < 4 - 4\omega\Lambda^2 + \omega^2\Lambda^4 = (2 - \omega\Lambda^2)^2.$$

Therefore, the spectral radius described by (4.11) is strictly monotonically decreasing for $0 < \omega < \omega_0$ and strictly monotonically increasing for $\omega_0 < \omega < 2$ (see Figure 4.1). Since $\rho[B(0)] = \rho[B(2)] = 1$, we finally obtain that $\rho[B(\omega)] < 1$ for all $0 < \omega < 2$ and that $\rho[B(\omega)]$ assumes its minimum for $\omega = \omega_0(\Lambda)$ with value $\rho[B(\omega_0(\Lambda))] = \omega_0(\Lambda) - 1$. □

Corollary 4.16 *Under the assumptions of Theorem 4.15 the Gauss–Seidel method converges twice as fast as the Jacobi method.*

Proof. From (4.8) we observe that $\mu = \lambda^2$ for $\omega = 1$; i.e., we have

$$\rho[B(1)] = \{\rho[-D^{-1}(A_L + A_R)]\}^2$$

for the spectral radii of the Gauss–Seidel matrix $B(1)$ and the Jacobi matrix $-D^{-1}(A_L + A_R)$. Now the statement follows from the observation that by the a priori estimate of Theorem 3.48 the number N of iterations required for a desired accuracy is inversely proportional to the modulus of the logarithm of the spectral radius; i.e.,

$$\frac{N(\text{Gauss–Seidel})}{N(\text{Jacobi})} \approx \frac{\ln \rho[-D^{-1}(A_L + A_R)]}{\ln \rho[B(1)]} = \frac{1}{2},$$

and this proves the assertion. □

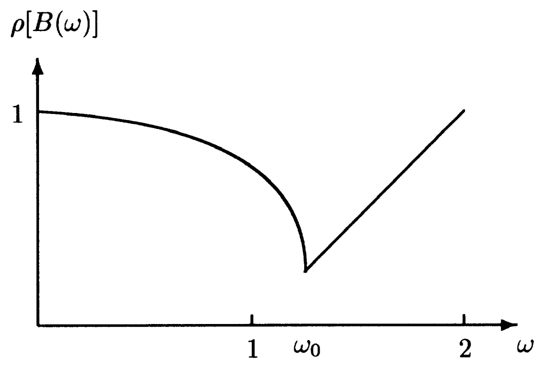


FIGURE 4.1. Spectral radius for SOR

Example 4.17 *For the tridiagonal matrix A from Example 4.5 we have*

$$\frac{N(\text{SOR})}{N(\text{Jacobi})} \approx \frac{\pi}{4(n+1)}$$

for the optimal relaxation parameter.

Proof. Using the trigonometric addition theorem

$$\frac{1}{2}\sin\frac{\pi j(k-1)}{n+1} + \frac{1}{2}\sin\frac{\pi j(k+1)}{n+1} = \cos\frac{\pi j}{n+1}\sin\frac{\pi jk}{n+1},$$

it can be seen that the Jacobi matrix

$$-D^{-1}(A_L + A_R) = \frac{1}{2}\begin{pmatrix} 0 & 1 & & & & \\ 1 & 0 & 1 & & & \\ & 1 & 0 & 1 & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & 1 & 0 & 1 \\ & & & & 1 & 0 \end{pmatrix}$$

corresponding to Example 4.5 has the eigenvalues

$$\lambda_j = \cos\frac{\pi j}{n+1}, \quad j = 1, \ldots, n,$$

and associated eigenvectors v_j with components

$$v_{j,k} = \sin\frac{\pi jk}{n+1}, \quad k = 1, \ldots, n, \quad j = 1, \ldots, n.$$

Hence,

$$\Lambda = \rho[-D^{-1}(A_L + A_R)] = \cos\frac{\pi}{n+1} \approx 1 - \frac{\pi^2}{2(n+1)^2}$$

and

$$-\ln \rho[-D^{-1}(A_L + A_R)] \approx \frac{\pi^2}{2(n+1)^2} .$$

From Theorem 4.15 we obtain

$$\omega_{\text{opt}} = \frac{2}{1 + \sin \dfrac{\pi}{n+1}}$$

and

$$\rho[B(\omega_{\text{opt}})] = \frac{1 - \sin \dfrac{\pi}{n+1}}{1 + \sin \dfrac{\pi}{n+1}} \approx 1 - \frac{2\pi}{n+1} ,$$

whence

$$-\ln \rho[B(\omega_{\text{opt}})] \approx \frac{2\pi}{n+1}$$

follows. This concludes the proof. □

For example, for $n = 30$ the optimal SOR method is about forty times as fast as the Jacobi method. Note that the improvement on the speed of convergence improves as n increases. The fact that in Example 4.17, and, more generally, in almost all linear systems arising in the discretization of boundary value problems, the optimal relaxation parameter has the property $\omega > 1$ explains why the method is known as the overrelaxation method.

4.3 Two-Grid Methods

Consider the linear system

$$Ax = y \tag{4.12}$$

with a nonsingular matrix A, and assume that we already have an approximate solution x_0 available with a *residual*, or *defect*,

$$r_0 := y - Ax_0,$$

for which, in general, $r_0 \neq 0$. Then we try to improve on the accuracy by writing

$$x_1 = x_0 + \delta_0 \tag{4.13}$$

with some correction term δ_0. Substituting this into (4.12) we obtain that δ_0 has to satisfy the defect correction equation

$$A\delta_0 = r_0$$

in order that x_1 satisfy (4.12). We observe that the correction term δ_0 will, in general, be small compared to x_0, and therefore it is unnecessary to solve the defect correction equation exactly. Hence we write

$$\delta_0 = A_{\text{approx}}^{-1} r_0,$$

where A_{approx}^{-1} is some approximation for the inverse A^{-1} of A. Substituting this into (4.13) we obtain

$$x_1 = x_0 + A_{\text{approx}}^{-1}[y - Ax_0] = (I - A_{\text{approx}}^{-1}A)x_0 + A_{\text{approx}}^{-1}y \tag{4.14}$$

as our new approximate solution to (4.12). This procedure is known as the *defect correction principle.*

Repeating this process yields the *defect correction iteration* defined by

$$x_{\nu+1} := x_\nu + A_{\text{approx}}^{-1}[y - Ax_\nu], \quad \nu = 0, 1, 2, \dots, \tag{4.15}$$

for the solution of (4.12). By Theorem 4.1, the iteration (4.15) converges to the unique solution x of $A_{\text{approx}}^{-1}[y - Ax] = 0$, provided that the spectral radius of the iteration matrix $I - A_{\text{approx}}^{-1}A$ is less than one. Since the unique solution x of $Ax = y$ trivially satisfies $A_{\text{approx}}^{-1}[y - Ax] = 0$, we then have convergence of the scheme (4.15) to the unique solution of (4.12). For a rapid convergence it is desirable that the spectral radius be close to zero, which will be the case if A_{approx}^{-1} is a reasonable approximation to A^{-1}. For a more complete introduction to the defect correction principle we refer to [56].

Here we wish to indicate briefly two applications. Firstly, the defect correction principle (4.14) can be used to improve on the accuracy of an approximate solution x_0, obtained for example by Gaussian elimination. Then, in principle, the computation of x_0 corresponds to some approximation $x_0 = A_{\text{approx}}^{-1}y$ obtained from an LR decomposition. This means that evaluating $\delta_0 = A_{\text{approx}}^{-1}r_0$ is achieved by applying again the same elimination algorithm to the defect correction equation. This way, the defect correction principle provides a simple tool to improve on the accuracy of a solution to a linear system obtained by elimination.

Secondly, we would like to illustrate the more systematic use of the defect correction principle for the development of multigrid methods as a powerful tool for the fast iterative solution of linear systems arising in the discretization of differential and integral equations. For the sake of simplicity we will confine ourselves to the case of two-grid iterations.

The basic idea of two-grid methods is to use the defect correction principle with the approximate inverse A_{approx}^{-1} for the matrix A_{fine} of a large linear system corresponding to a fine approximation grid given simply by the exact inverse of the matrix A_{coarse} of a smaller linear system, corresponding to a coarse approximation grid. Of course, a number of mathematical problems arise in the design of such methods concerning the appropriate

relation between the fine and coarse grid and the transfer between the two grids. We will outline some ideas on the structure of two-grid methods by again considering the simple model problem from Example 2.1 as a typical case.

Recall that the solution vector $U^{(h)} \in \mathbb{R}^n$ of the linear system

$$A^{(h)}U^{(h)} = F^{(h)} \tag{4.16}$$

with the $n \times n$ tridiagonal matrix

$$A^{(h)} = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}$$

corresponds to approximate values $u_j^{(h)} \approx u(jh)$, $j = 1, \dots, n$, for the solution u of the boundary value problem (2.1)–(2.2) at the internal grid points. Since we want to make use of two different grids in our analysis, we indicate the dependence on the mesh width

$$h = \frac{1}{n+1}$$

in the matrix $A^{(h)}$ and the solution $U^{(h)}$. We assume that n is odd because later we want to choose the coarser grid by doubling the mesh width.

We start from the Jacobi iteration with relaxation

$$U_{\nu+1}^{(h)} = U_\nu^{(h)} - \omega[D^{(h)}]^{-1}[A^{(h)}U_\nu^{(h)} - F^{(h)}], \quad \nu = 0, 1, 2, \dots, \tag{4.17}$$

as introduced in Definition 4.8. From our analysis in Example 4.17 we deduce that $A^{(h)}$ has the n eigenvalues

$$\mu_j = \frac{4}{h^2}\sin^2\frac{\pi jh}{2}, \quad j = 1, \dots, n, \tag{4.18}$$

and associated eigenvectors $v_j^{(h)}$ with components

$$v_{j,k}^{(h)} = \sin(\pi jkh), \quad k = 1, \dots, n, \quad j = 1, \dots, n. \tag{4.19}$$

Note that by Theorem 3.29, the eigenvectors of the Hermitian matrix $A^{(h)}$ form an orthogonal basis for $\mathbb{R}^n$ (see Problem 4.18). The $v_j^{(h)}$, $j = 1, \dots, n$, are also eigenvectors of the Jacobi matrix $I - [D^{(h)}]^{-1}A^{(h)}$, with eigenvalues

$$\lambda_j = \cos(\pi jh), \quad j = 1, \dots, n.$$

From Theorem 4.9 we observe that $\omega = 1$ is the optimal choice for the Jacobi iteration with relaxation. However, it will turn out that in the context of two-grid methods the *damped*, or *underrelaxed, Jacobi method* with $0 < \omega < 1$ is more important. This is due to the following observation. Since the $v_j^{(h)}$, $j = 1, \dots, n$, provide a basis for $\mathbb{R}^n$, we can represent the difference between the exact solution $U^{(h)}$ and the νth iteration U_ν in the form

$$U^{(h)} - U_\nu = \sum_{j=1}^{n} \alpha_{j,\nu} v_j^{(h)}.$$

From the fact that

$$\{I - \omega [D^{(h)}]^{-1} A^{(h)}\} v_j^{(h)} = \left\{ 1 - 2\omega \sin^2 \frac{\pi j h}{2} \right\} v_j^{(h)}, \quad j = 1, \dots, n,$$

we derive the recurrence relation

$$\alpha_{j,\nu+1} = \left\{ 1 - 2\omega \sin^2 \frac{\pi j h}{2} \right\} \alpha_{j,\nu}, \quad j = 1, \dots, n,$$

for the coefficients $\alpha_{j,\nu}$. In particular, if we choose $\omega = 0.5$, we have that

$$\alpha_{j,\nu+1} = \cos^2 \frac{\pi j h}{2} \, \alpha_{j,\nu}, \quad j = 1, \dots, n. \tag{4.20}$$

From this we observe that even though convergence of the iterations (4.17) becomes slower when we decrease ω, for $\omega = 0.5$ the convergence restricted to the subspace

$$W_n := \operatorname{span}\{v_{\frac{n+1}{2}}, \dots, v_n\}$$

of *high frequencies* is dramatically accelerated, since in this case from (4.20) we have that

$$|\alpha_{j,\nu+1}| \le \frac{1}{2} \, |\alpha_{j,\nu}|, \quad j = \frac{n+1}{2}, \dots, n.$$

This fact can be expressed by saying that the damped Jacobi iteration is a *smoothing iteration.* In the sequel we will consider only the damping factor $\omega = 0.5$.

The slow convergence with respect to low frequencies will now be taken care of by the defect correction principle through incorporating a so-called *coarse grid correction* on the grid with mesh width $2h$. For this we need to transfer vectors corresponding to the fine grid to vectors corresponding to the coarse grid and vice versa. The transfer from the fine grid to the coarse grid requires a *restriction* and corresponds to a mapping $R^{(h)} : \mathbb{R}^n \to \mathbb{R}^{\frac{n-1}{2}}$. Note that we only need to consider this mapping for the interior grid points. Instead of choosing the restriction $(R^{(h)} y)_k = y_{2k}$,

$k = 1, \ldots, \frac{n-1}{2}$, for $y \in \mathbb{R}^n$ it turns out to be advantageous to also incorporate information contained in the odd nodal points of the fine grid by using the restriction

$$(R^{(h)} y)_k = \frac{1}{4}\,[y_{2k-1} + 2y_{2k} + y_{2k+1}], \quad k = 1, \ldots, \frac{n-1}{2}\,,$$

as illustrated in Figure 4.2.

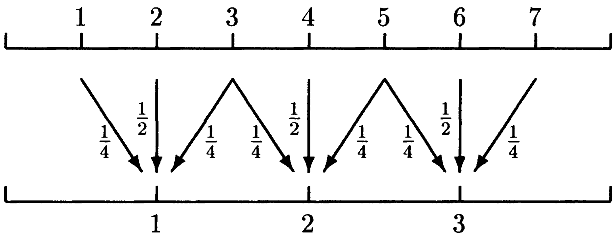


FIGURE 4.2. Restriction operator of the two-grid method for $n = 7$

The corresponding matrix is

$$R^{(h)} = \frac{1}{4}\begin{pmatrix} 1 & 2 & 1 & & & & & & & & \\ & & 1 & 2 & 1 & & & & & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & & & 1 & 2 & 1 & & \\ & & & & & & & & 1 & 2 & 1 \end{pmatrix}.$$

With the aid of elementary trigonometric manipulations one can establish the relation

$$R^{(h)} v_j^{(h)} = c_j^2 v_j^{(2h)}, \quad R^{(h)} v_{n+1-j}^{(h)} = -s_j^2 v_j^{(2h)}, \quad j = 1, \ldots, \frac{n-1}{2}\,, \tag{4.21}$$

between the eigenvectors (4.19) for the fine and the coarse grid (see Problem 4.19). Here we have set

$$c_j = \cos\frac{j\pi h}{2}\,, \quad s_j = \sin\frac{j\pi h}{2}\,, \quad j = 1, \ldots, \frac{n-1}{2}\,.$$

The transfer from the coarse grid to the fine grid is called *prolongation* and corresponds to a mapping $P^{(h)} : \mathbb{R}^{\frac{n-1}{2}} \to \mathbb{R}^n$. The simplest choice for $P^{(h)}$ is given by the piecewise linear interpolation (see Chapter 8)

$$(P^{(h)} y)_{2k} = y_k, \qquad k = 1, \ldots, \frac{n-1}{2}\,,$$

$$(P^{(h)} y)_{2k-1} = \frac{1}{2}\,[y_k + y_{k-1}], \qquad k = 1, \ldots, \frac{n+1}{2}\,,$$

for $y \in \mathbb{R}^{\frac{n-1}{2}}$, as illustrated in Figure 4.3. The corresponding matrix is given by $P^{(h)} = 2R^{(h)\perp}$. Either by direct computation or from (4.21) and the fact that the matrices $P^{(h)}$ and $2R^{(h)}$ are adjoint one can establish that (see Problem 4.19)

$$P^{(h)}v_j^{(2h)} = c_j^2 v_j^{(h)} - s_j^2 v_{n+1-j}^{(h)}, \quad j = 1, \dots, \frac{n-1}{2}. \tag{4.22}$$

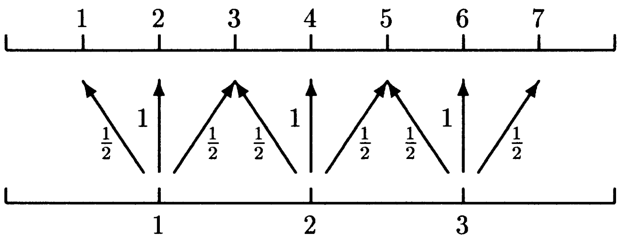


FIGURE 4.3. Prolongation operator of the two-grid method for $n = 7$

Now we are in a position to use the $n \times n$ matrix $P^{(h)}[A^{(2h)}]^{-1}R^{(h)}$ as the coarse-grid correction. Computing $P^{(h)}[A^{(2h)}]^{-1}R^{(h)}y$ corresponds to first restricting the vector $y \in \mathbb{R}^n$ to $R^{(h)}y \in \mathbb{R}^{\frac{n-1}{2}}$, then solving the $\frac{n-1}{2} \times \frac{n-1}{2}$ system $A^{(2h)}z = R^{(h)}y$ by an elimination method, and finally prolonging the solution $z \in \mathbb{R}^{\frac{n-1}{2}}$ to $P^{(h)}z \in \mathbb{R}^n$. Combining this coarse-grid correction with N steps of the damped Jacobi iteration in the sense of (4.14) now yields one step of the two-grid iteration scheme

$$U_{\nu+1} = J_N(U_\nu, F^{(h)}) - P^{(h)}[A^{(2h)}]^{-1}R^{(h)}[A^{(h)}J_N(U_\nu, F^{(h)}) - F^{(h)}],$$

where $J_N(U_\nu, F^{(h)})$ denotes the result of N steps of the damped Jacobi iterations (4.17) with starting element U_ν. Obviously, the iteration matrix corresponding to this two-grid method is given by

$$T_N = \{I - P^{(h)}[A^{(2h)}]^{-1}R^{(h)}A^{(h)}\}\left\{I - \frac{1}{2}\,[D^{(h)}]^{-1}A^{(h)}\right\}^N. \tag{4.23}$$

For an investigation of the convergence for our two-grid iteration scheme we need to determine the spectral radius of T_N. For simplicity we confine ourselves to the case where $N = 1$; i.e., one step of the damped Jacobi iteration on the fine grid alternates with a coarse-grid correction by elimination on the coarse grid. We set $T_1 = T$.

Theorem 4.18 *For the spectral radius of T we have that $\rho(T) = 0.5$; i.e., the two-grid iterations converge.*

Proof. We note that from (4.18) and (4.19), with h replaced by $2h$, we have that

$$A^{(2h)}v_j^{(2h)} = \frac{1}{h^2}\sin^2(\pi jh)\,v_j^{(2h)} = \frac{4}{h^2}\,c_j^2 s_j^2\,v_j^{(2h)},$$

whence

$$[A^{(2h)}]^{-1}v_j^{(2h)} = \frac{h^2}{4c_j^2 s_j^2}\,v_j^{(2h)}, \quad j = 1,\ldots,\frac{n-1}{2}\,,$$

follows. From this, using (4.20)–(4.22) and $R^{(h)}v^{(h)}_{\frac{n+1}{2}} = 0$, it can be derived that

$$\begin{pmatrix} Tv_j^{(h)} \\ Tv_{n+1-j}^{(h)} \end{pmatrix} = s_j^2 c_j^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} v_j^{(h)} \\ v_{n+1-j}^{(h)} \end{pmatrix} \tag{4.24}$$

for $j = 1,\ldots,\frac{n-1}{2}$ and

$$Tv^{(h)}_{\frac{n+1}{2}} = \frac{1}{2}\,v^{(h)}_{\frac{n+1}{2}}\,. \tag{4.25}$$

Since the matrix

$$Q = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

has the eigenvalues 0 and 2, from (4.24) and (4.25) it can be seen that the matrix T has the eigenvalues

$$2s_j^2c_j^2 = \frac{1}{2}\sin^2\pi jh, \quad j = 1,\ldots,\frac{n+1}{2}\,,$$

and the eigenvalue zero of multiplicity $\frac{n-1}{2}$. This implies the assertion on the spectral radius of T. □

Theorem 4.18 shows that the two-grid method is a very fast iteration. As compared to the classical Jacobi and Gauss–Seidel methods and also to the SOR method with optimal relaxation parameter, it decreases the spectral radius from a value close to one to one-half, which causes a substantial increase in the speed of convergence. However, for practical computations it has the disadvantage that in each step the solution of a system with half the number of unknows is required.

This drawback of the two-grid method is remedied by the multigrid method. Whereas for the two-grid method as described above only two grids are used, the multigrid method uses $M > 2$ different grids with mesh widths $h_\mu = 2^\mu h, \mu = 1,\ldots,M$, obtained from the mesh width h on the finest grid. The *multigrid method* is defined recursively. The method for $M+1$ grids performs one or several steps of the damped Jacobi iteration on the finest grid with mesh width h and uses as approximate inverse for the defect correction one or several steps of the multigrid iteration on the M grids with mesh widths $2h, 4h, \ldots, 2^M h$. To be more explicit, the three-grid method uses one or several steps of the two-grid method as the defect

correction of the damped Jacobi iteration on the finest grid; the four-grid method uses one or several steps of the three-grid method as the defect correction; and so on. To describe further details of the multigrid method, in particular showing that the computational cost of one step of a multigrid iteration is proportional to the cost of the Jacobi iterations on the finest grid provided that the coarsest grid is coarse enough, is beyond the aim of this introduction. For a comprehensive study we refer to [8, 26, 29, 63].

Problems

4.1 Consider the solution of the linear system

$$\begin{aligned} 5x_1 \qquad\quad - 2x_3 &= -1 \\ -4x_1 + 8x_2 + 2x_3 &= 18 \\ 5x_2 + 9x_3 &= 37 \end{aligned}$$

by the Jacobi method. Give an estimate on the number of iterations needed to ensure that $\|x_\nu - x\|_\infty \le 10^{-3}$ if the iteration is started with $x_0 = (0,0,0)^T$.

4.2 Write a computer program for the Jacobi method, the Gauss–Seidel method, and the SOR method and test it for various examples.

4.3 Show that a matrix A has spectral radius $\rho(A) < 1$ if and only if it satisfies $\lim_{\nu\to\infty} A^\nu = 0$.

4.4 Prove that the Jacobi method converges for strictly column-diagonally dominant matrices (compare (4.5)).

4.5 Show that an $n \times n$ matrix A is reducible if and only if there exists an $n \times n$ permutation matrix P such that

$$P^{-1}AP = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix},$$

where A_{11} is a $k\times k$ matrix and A_{22} is an $(n-k)\times(n-k)$ matrix with $1 \le k \le n-1$.

4.6 Show that the matrix A from Example 4.5 is irreducible and weakly row-diagonally dominant.

4.7 Let

$$A = \begin{pmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{pmatrix}.$$

Show that for $1 \le 2\alpha < 2$ the Gauss–Seidel method is convergent and the Jacobi method is not.

4.8 For the matrix

$$A = \begin{pmatrix} 1 & -2 & 2 \\ -1 & 1 & -1 \\ -2 & -2 & 1 \end{pmatrix}$$

show that the Jacobi method is convergent and the Gauss–Seidel method is not.

4.9 For the matrix

$$A = \begin{pmatrix} 2 & 1 & -1 \\ -2 & 2 & -2 \\ 1 & 1 & 2 \end{pmatrix}$$

show that the Gauss–Seidel method is convergent and the Jacobi method is not.

4.10 Show that the matrix

$$A = \begin{pmatrix} 2 & 0 & -1 & -1 \\ 0 & 2 & -1 & -1 \\ -1 & -1 & 2 & 0 \\ -1 & -1 & 0 & 2 \end{pmatrix}$$

is irreducible and that the Jacobi method is not convergent.

4.11 Show that the iteration matrix of the Gauss–Seidel method has eigenvalue zero.

4.12 Consider the variant of the Gauss–Seidel iteration where the components are iterated from the nth component backward to the first component. What is the iteration matrix of this method? Obtain a symmetric method by alternating one step of the forward Gauss–Seidel method and one step of the backward Gauss–Seidel method. What is the iteration matrix of this method?

4.13 Show that the Jacobi iteration converges for a matrix A if and only if it converges for the transposed matrix A^T.

4.14 Show that the matrix A of Example 2.2 is irreducible, positive definite, and weakly row-diagonally dominant.

4.15 Compute the eigenvalues of the Jacobi iteration matrix for the matrix A of Example 2.2.

4.16 Let $A = (a_{jk})$ be a nonnegative $n \times n$ matrix, i.e., $a_{jk} \geq 0$, $j, k = 1, \ldots, n$, and let $\rho(A) < 1$. Show that $I - A$ is nonsingular and $(I - A)^{-1}$ is nonnegative.

4.17 Give a counterexample to show that the Jacobi method, in general, does not converge for positive definite matrices (see Theorem 4.12).

4.18 Show by direct computations that the eigenvectors given by (4.19) are orthogonal.

4.19 Prove the relations (4.21), (4.22), (4.24), and (4.25).

4.20 Show that

$$\rho(T_N) \leq \max_{0 \leq t \leq \frac{1}{2}} \left[t(1-t)^N + (1-t)^N t\right]$$

for the two-grid iteration matrix with N damped Jacobi iterations at each step.

5

Ill-Conditioned Linear Systems

For problems in mathematical physics Hadamard [31] postulated three requirements: A solution should exist, the solution should be unique, and the solution should depend continuously on the data. The third postulate is motivated by the fact that in general, in applications the data will be measured quantities and therefore always contaminated by errors. A problem satisfying all three requirements is called *well-posed.* Otherwise, it is called *ill-posed.* If $A : X \to Y$ is a bounded linear operator mapping a normed space X into a normed space Y, then the equation $Ax = y$ is well-posed if A is bijective and the inverse operator $A^{-1} : Y \to X$ is bounded (see Theorem 3.24). Since the inverse of a linear operator again is linear, in the case of finite-dimensional spaces X and Y, by Theorem 3.26 bijectivity of A implies boundedness of the inverse operator. Hence, in the sense of Hadamard, nonsingular linear systems are well-posed.

However, since one wants to make sure that small errors in the data of a linear system will cause only small errors in the solution, there is an additional need for a measure of the degree of well-posedness, or stability. Such a measure is provided through the notion of the condition number, which we will introduce in this chapter. This will enable us to distinguish between *well-conditioned* and *ill-conditioned* linear systems. For the latter, small errors in the data may cause large errors in the solution, and therefore their numerical solution requires special care.

Hence, we will continue the chapter with a brief discussion of the singular value cutoff and the Tikhonov regularization as efficient means to deal with ill-conditioned linear systems. Our analysis will be based on the singular value decomposition and will include the introduction of the pseudo-inverse,

or Moore–Penrose inverse. For an extension of these ideas to ill-posed linear operator equations in infinite-dimensional spaces we refer to [14, 22, 28, 37, 39, 43].

5.1 Condition Number

We begin with an example of an ill-conditioned linear system arising through a simple least squares problem.

Example 5.1 We consider the best approximation of a given continuous function $f : [0,1] \to \mathbb{R}$ by a polynomial

$$p(x) = \sum_{k=0}^{n} \alpha_k x^k$$

of degree n in the least squares sense, i.e., with respect to the L_2 norm. Using the monomials $x \mapsto x^k$, $k = 0, 1, \dots, n$, as a basis of the subspace $P_n \subset C[0,1]$ of polynomials of degree less than or equal to n (see Theorem 8.2), from Corollary 3.53 and the integrals

$$\int_0^1 x^j x^k \, dx = \frac{1}{j+k+1}$$

it follows that the coefficients $\alpha_0, \dots, \alpha_n$ of the best approximation are uniquely determined by the normal equations

$$\sum_{k=0}^{n} \frac{1}{j+k+1} \, \alpha_k = \int_0^1 f(x) x^j \, dx, \quad j = 0, \dots, n. \tag{5.1}$$

In the special case

$$f(x) = \frac{1}{1+x}$$

we have the right-hand sides

$$r_j := \int_0^1 \frac{x^j}{1+x} \, dx, \quad j = 0, \dots, n.$$

In particular, $r_0 = \ln 2$, and from the geometric sum

$$\sum_{i=1}^{j} (-1)^{i-1} x^{i-1} = \frac{1 - (-1)^j x^j}{1+x}, \quad j = 1, \dots, n,$$

we deduce that

$$r_j = (-1)^j \left\{ \ln 2 + \sum_{i=1}^{j} (-1)^i \frac{1}{i} \right\}, \quad j = 1, \dots, n.$$

Therefore, the solution of (5.1) is of the form

$$\alpha_j = \beta_j \ln 2 + \gamma_j, \quad j = 0, \dots, n,$$

with rational numbers β_j and γ_j. Table 5.1 gives the exact solution of the linear system (5.1) obtained by Gaussian elimination carried out in terms of rational numbers to compute the coefficients β_j and γ_j and then inserting $\ln 2$ with ten-decimal-digit accuracy. The results indicate convergence of the coefficients to the coefficients $a_k = (-1)^k$ of the Taylor series for f.

TABLE 5.1. Exact solution of the linear system (5.1)

n	α_0	α_1	α_2	α_3	α_4	α_5	α_6
1	0.9314	−0.4766					
2	0.9860	−0.8040	0.3274				
3	0.9972	−0.9389	0.6645	−0.2247			
4	0.9994	−0.9830	0.8630	−0.5334	0.1543		
5	0.9999	−0.9956	0.9512	−0.7688	0.4191	−0.1059	
6	0.9999	−0.9989	0.9843	−0.9011	0.6672	−0.3242	0.0727

However, if we take as right-hand sides the values obtained for r_j by using $\ln 2$ with five-decimal-digit accuracy, then Gaussian elimination yields the results of Table 5.2.

TABLE 5.2. Numerical solution of the linear system (5.1)

n	α_0	α_1	α_2	α_3	α_4	α_5	α_6
1	0.93	−0.47					
2	0.98	−0.80	0.32				
3	0.99	−0.95	0.70	−0.24			
4	1.00	−1.16	1.63	−1.69	0.72		
5	1.06	−2.74	12.68	−31.16	33.87	−13.25	
6	1.39	−16.58	151.09	−584.79	1071.93	−926.75	304.49

Despite the fact that the changes in the right-hand sides are less than 0.000005, we obtain drastic changes in the solution. Therefore, qualitatively we may say that our linear system provides an example of an ill-conditioned system. The matrix of this example is known as the *Hilbert matrix.* □

For a quantitative analysis of the phenomenon illustrated by Example 5.1 we introduce the concept of the condition number.

Definition 5.2 *Let X and Y be normed spaces and let $A : X \to Y$ be a bounded linear operator with a bounded inverse $A^{-1} : Y \to X$. Then*

$$\operatorname{cond}(A) := \|A\| \, \|A^{-1}\|$$

is called the condition number *of A.*

Clearly, $\operatorname{cond}(A)$ depends on the chosen norm. Because of (see Remark 3.25)

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\| \, \|A^{-1}\|$$

we always have $\operatorname{cond}(A) \geq 1$. Definition 5.2, in particular, includes the condition number of a nonsingular $n \times n$ matrix A. Here, in the case where both the domain and range are given the ℓ_p norm for $p = 1, 2, \infty$ we will write $\operatorname{cond}_p(A)$.

Theorem 5.3 *Let X and Y be Banach spaces, let $A : X \to Y$ be a bounded linear operator with a bounded inverse $A^{-1} : Y \to X$ and let $A^\delta : X \to Y$ be a bounded linear operator such that $\|A^{-1}\| \, \|A^\delta - A\| < 1$. Assume that x and x^δ are solutions of the equations*

$$Ax = y \tag{5.2}$$

and

$$A^\delta x^\delta = y^\delta, \tag{5.3}$$

respectively. Then

$$\frac{\|x^\delta - x\|}{\|x\|} \leq \frac{\operatorname{cond}(A)}{1 - \operatorname{cond}(A) \dfrac{\|A^\delta - A\|}{\|A\|}} \left\{ \frac{\|y^\delta - y\|}{\|y\|} + \frac{\|A^\delta - A\|}{\|A\|} \cdot \right\}$$

Proof. Writing $A^\delta = A[I + A^{-1}(A^\delta - A)]$, by Theorem 3.48 we observe that the inverse operator $[A^\delta]^{-1} = [I + A^{-1}(A^\delta - A)]^{-1}A^{-1}$ exists and is bounded by

$$\|[A^\delta]^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \, \|A^\delta - A\|} \, . \tag{5.4}$$

From (5.2) and (5.3) we find that

$$A^\delta(x^\delta - x) = y^\delta - y - (A^\delta - A)x,$$

whence

$$x^\delta - x = [A^\delta]^{-1}\{y^\delta - y - (A^\delta - A)x\}$$

follows. Now we can estimate

$$\|x^\delta - x\| \leq \|[A^\delta]^{-1}\|\{\|y^\delta - y\| + \|A^\delta - A\| \, \|x\|\}$$

and insert (5.4) to obtain

$$\frac{\|x^\delta - x\|}{\|x\|} \leq \frac{\operatorname{cond}(A)}{1 - \|A^{-1}\|\,\|A^\delta - A\|}\left\{\frac{\|y^\delta - y\|}{\|A\|\,\|x\|} + \frac{\|A^\delta - A\|}{\|A\|}\right\}.$$

From this the assertion follows with the aid of $\|A\|\,\|x\| \geq \|y\|$. □

Theorem 5.3 shows that the condition number may serve as a measure of stability for linear operator equations and, in particular, for linear systems. A linear system with a small condition number is stable, whereas a large condition number indicates instability. We call a linear system with a small condition number *well-conditioned.* Otherwise, it is called *ill-conditioned.*

By Theorem 3.31, the condition number of a Hermitian matrix A in the Euclidean norm is given by

$$\operatorname{cond}_2(A) = \frac{|\lambda_{\max}|}{|\lambda_{\min}|},$$

where $\lambda_{\max}$ and $\lambda_{\min}$ denote the eigenvalues of A with largest and smallest modulus, respectively. Table 5.3 is obtained by employing the QR algorithm (see Section 7.4) for the computation of matrix eigenvalues. It illustrates quantitatively the degree of instability, i.e., the ill-conditionedness of the linear system from Example 5.1.

TABLE 5.3. Condition number for the linear system (5.1)

n	2	3	4	5	6
$\lambda_{\max}$	1.27	1.41	1.50	1.57	1.62
$\lambda_{\min}$	$6.57 \cdot 10^{-2}$	$2.69 \cdot 10^{-3}$	$9.67 \cdot 10^{-5}$	$3.29 \cdot 10^{-6}$	$1.08 \cdot 10^{-7}$
cond_2	19.3	$5.24 \cdot 10^{2}$	$1.55 \cdot 10^{4}$	$4.77 \cdot 10^{5}$	$1.50 \cdot 10^{7}$

5.2 Singular Value Decomposition

In the sequel we wish to introduce some of the basic concepts for the approximate solution of ill-conditioned linear systems. Our approach will be based on the singular value decomposition of a matrix A, which need not be a square matrix.

For each $m \times n$ matrix A, representing an operator $A : \mathbb{C}^n \to \mathbb{C}^m$, the $n \times n$ matrix A^*A is Hermitian and positive semidefinite (see Problem 5.9). Therefore, the eigenvalues of A^*A are real and nonnegative (see Theorem 3.29). The nonnegative square roots of these eigenvalues are called the *singular values* of A.

For the remainder of this chapter, by $(\cdot\,,\cdot)$ we denote the Euclidean scalar product in $\mathbb{C}^n$. For an $m \times n$ matrix A of rank r, the nullspace

$N(A) = \{x \in \mathbb{C}^n : Ax = 0\}$ has dimension $\dim N(A) = n - r$. We note that $A^*Au = 0$ implies that

$$\|Au\|_2 = (Au, Au) = (u, A^*Au) = 0;$$

i.e., the nullspaces of A and A^*A coincide. Hence $\dim N(A^*A) = n - r$, and therefore A has exactly r positive singular values μ (counted according to their geometric multiplicity, i.e., according to the dimension of the nullspace of $\mu^2 I - A^*A$).

Theorem 5.4 *Let A be an $m \times n$ matrix of rank r. Then there exist nonnegative numbers*

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_r > \mu_{r+1} = \cdots = \mu_n = 0$$

and orthonormal vectors $u_1, \dots, u_n \in \mathbb{C}^n$ and $v_1, \dots, v_m \in \mathbb{C}^m$ such that

$$\begin{aligned} Au_j &= \mu_j v_j, \quad A^*v_j = \mu_j u_j, \quad j = 1, \dots, r, \\ Au_j &= 0, \quad j = r+1, \dots, n, \\ A^*v_j &= 0, \quad j = r+1, \dots, m. \end{aligned} \tag{5.5}$$

For each $x \in \mathbb{C}^n$ we have the singular value decomposition

$$Ax = \sum_{j=1}^{r} \mu_j (x, u_j)\, v_j. \tag{5.6}$$

Each system (μ_j, u_j, v_j) with these properties is called a singular system *of the matrix A.*

Proof. The Hermitian and semipositive definite matrix A^*A of rank r has n orthonormal eigenvectors $u_1, \dots, u_n$ with nonnegative eigenvalues

$$A^*Au_j = \mu_j^2 u_j, \quad j = 1, \dots, n, \tag{5.7}$$

which we may assume to be ordered according to $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_r > 0$ and $\mu_{r+1} = \cdots = \mu_n = 0$. We define

$$v_j := \frac{1}{\mu_j}\, Au_j, \quad j = 1, \dots, r.$$

Then, using (5.7) we have

$$(v_j, v_k) = \frac{1}{\mu_j \mu_k}\,(Au_j, Au_k) = \frac{1}{\mu_j \mu_k}\,(u_j, A^*Au_k) = \delta_{jk}, \quad j, k = 1, \dots, r,$$

where $\delta_{jk} = 1$ for $k = j$, and $\delta_{jk} = 0$ for $k \neq j$. Further, we compute that $A^*v_j = \mu_j u_j$, $j = 1, \dots, r$, and hence the first line of (5.5) is proven. The second line of (5.5) is a consequence of $N(A) = N(A^*A)$.

If $r < m$, by the Gram–Schmidt orthogonalization procedure from Theorem 3.18 we can extend $v_1, \dots, v_r$ to an orthonormal basis $v_1, \dots, v_m$ of $\mathbb{C}^m$. Since A^* has rank r, we have $\dim N(A^*) = m - r$. From this we can conclude the third line of (5.5).

Since the $u_1, \dots, u_n$ form an orthonormal basis of $\mathbb{C}^n$, we can represent

$$x = \sum_{j=1}^{n} (x, u_j)\, u_j,$$

and (5.6) follows by applying A and observing (5.5). □

Clearly, we can rewrite the equations (5.5) in the form

$$A = VDU^*, \tag{5.8}$$

where $U = (u_1, \dots, u_n)$ and $V = (v_1, \dots, v_m)$ are unitary $n \times n$ and $m \times m$ matrices, respectively, and where D is an $m \times n$ diagonal matrix with entries $d_{jj} = \mu_j$ for $j = 1, \dots, r$ and $d_{jk} = 0$ otherwise.

Theorem 5.5 *Let A be an $m \times n$ matrix of rank r with singular system (μ_j, u_j, v_j). The linear system*

$$Ax = y \tag{5.9}$$

is solvable if and only if

$$(y, z) = 0 \tag{5.10}$$

for all $z \in \mathbb{C}^m$ with $A^* z = 0$. In this case a solution of (5.9) *is given by*

$$x_0 = \sum_{j=1}^{r} \frac{1}{\mu_j}\, (y, v_j)\, u_j. \tag{5.11}$$

Proof. Let x be a solution of (5.9) and let $A^* z = 0$. Then

$$(y, z) = (Ax, z) = (x, A^* z) = 0.$$

This implies the necessity of condition (5.10) for the solvability of (5.9).

Conversely, assume that (5.10) is satisfied. In terms of the orthonormal basis $v_1, \dots, v_m$ of $\mathbb{C}^m$ condition (5.10) implies that

$$y = \sum_{j=1}^{r} (y, v_j)\, v_j, \tag{5.12}$$

since $A^* v_j = 0$ for $j = r + 1, \dots, m$. For the vector x_0 defined by (5.11) we have that

$$Ax_0 = \sum_{j=1}^{r} (y, v_j)\, v_j.$$

In view of (5.12) this implies that $Ax_0 = y$, and the proof is complete. □

Since $N(A) = \text{span}\{u_{r+1}, \dots, u_n\}$, the vector x_0 defined by (5.11) has the property

$$(x_0, x) = 0$$

for all $x \in N(A)$. In the case where equation (5.9) has more than one solution, the general solution is obtained from (5.11) by adding an arbitrary solution x of the homogeneous equation $Ax = 0$. Then from

$$\|x_0 + x\|_2^2 = \|x_0\|_2^2 + 2\,\text{Re}(x_0, x) + \|x\|_2^2 = \|x_0\|_2^2 + \|x\|_2^2$$

we observe that (5.11) represents the uniquely determined solution of (5.9) with minimal Euclidean norm.

In the case where equation (5.9) has no solution, we represent

$$y = \sum_{j=1}^{m} (y, v_j)\, v_j$$

in terms of the orthonormal basis $v_1, \dots, v_m$. Let x_0 be given by (5.11) and let $x \in \mathbb{C}^n$ be arbitrary. Then

$$(Ax - Ax_0, Ax_0 - y) = 0,$$

since $Ax - Ax_0 \in \text{span}\{v_1, \dots, v_r\}$ and $Ax_0 - y \in \text{span}\{v_{r+1}, \dots, v_m\}$. This implies

$$\|Ax - y\|_2^2 = \|Ax - Ax_0\|_2^2 + \|Ax_0 - y\|_2^2,$$

whence (5.11) represents a least squares solution of (5.9) (see Example 2.4). Again, it can be shown that (5.11) is the uniquely determined least squares solution of (5.9) with minimal Euclidean norm (see Problem 5.11).

Hence, (5.11) defines a linear operator $A^\dagger : \mathbb{C}^m \to \mathbb{C}^n$ by

$$A^\dagger y := \sum_{j=1}^{r} \frac{1}{\mu_j}\, (y, v_j)\, u_j, \quad y \in \mathbb{C}^m, \tag{5.13}$$

which of course also allows a representation by an $n \times m$ matrix. Due to the properties of $A^\dagger y$ as discussed above, this operator or matrix is known as the *pseudo-inverse* or *Moore–Penrose inverse* of A (see [7]). It was first introduced by Moore in 1920 and independently rediscovered by Penrose in 1955. For an alternative introduction of $A^\dagger$ see Problem 5.12.

By Theorem 3.31 the condition number of a nonsingular matrix with respect to the Euclidean norm is given by the quotient of the largest and smallest singular value. Theorem 5.5 demonstrates the influence of small singular values on the condition of the matrix A. If for some $\delta \in \mathbb{C}$ we perturb the right-hand side by setting $y^\delta = y + \delta v_j$, we obtain a perturbed solution $x^\delta = x + \delta u_j/\mu_j$. Hence, the ratio $\|x^\delta - x\|_2 / \|y^\delta - y\|_2 = 1/\mu_j$ becomes large if A possesses small singular values.

This observation suggests stabilizing an ill-conditioned linear system by damping or filtering out the influence of the factor $1/\mu_j$ in the solution formula (5.11). In the so-called *spectral cutoff*, the terms in (5.11) corresponding to small singular values are simply neglected. Of course, this requires some strategy on how to determine the number of terms being summed up in (5.11). A very effective strategy is provided by the following *discrepancy principle.* If the right-hand side y of a linear system is known only within an error level δ then it is quite natural to require $Ax = y$ to be satisfied only up to the same accuracy δ, since it does not make much sense to try to satisfy the linear system more accurately than the right-hand side is known. To describe the discrepancy principle more precisely, given an erroneous right-hand side y^δ with known error level $\|y^\delta - y\|_2 \le \delta$, in the spectral cutoff the solution $x = A^\dagger y$ of $Ax = y$ is approximated by

$$x_p := \sum_{j=1}^{p} \frac{1}{\mu_j}\,(y^\delta, v_j)\,u_j \tag{5.14}$$

for some $0 \le p \le r$. For the following theorem we have to assume that $Ax = y$ is solvable.

Theorem 5.6 *Let A be an $m \times n$ matrix with singular system (μ_j, u_j, v_j) and let $y \in A(\mathbb{C}^n)$, $y^\delta \in \mathbb{C}^m$ satisfy*

$$\|y^\delta - y\|_2 \le \delta \le \|y^\delta\|_2$$

for $\delta > 0$. Then there exists a smallest integer $p = p(\delta)$ such that

$$\|Ax_p - y^\delta\|_2 \le \delta. \tag{5.15}$$

This discrecancy principle for the spectral cutoff is regular in the sense that if the error level δ tends to zero, then

$$x_p \to A^\dagger y, \quad \delta \to 0. \tag{5.16}$$

Proof. Consider the function $F : \{0, 1, \dots, r\} \to \mathbb{R}$ defined by

$$F(p) := \|Ax_p - y^\delta\|_2^2 - \delta^2.$$

In terms of the singular system, we can write

$$F(p) = \sum_{j=p+1}^{m} |(y^\delta, v_j)|^2 - \delta^2. \tag{5.17}$$

Hence, F is monotonically nonincreasing with $F(0) = \|y^\delta\|^2 - \delta^2 \ge 0$ and $F(r) = -\delta^2 < 0$ if the rank r of A is equal to m. If $r < m$, then using $(y, v_j) = 0$, $j = r+1, \dots, m$ (see the proof of Theorem 5.5), we have

$$F(r) = \sum_{j=r+1}^{m} |(y^\delta - y, v_j)|^2 - \delta^2 \le \|y^\delta - y\|_2^2 - \delta^2 \le 0.$$

Therefore, there exists a smallest integer $p = p(\delta)$ such that $F(p) \leq 0$. Note that $p \leq r$. In actual computations, this stopping parameter p is determined by terminating the sum (5.14) when the right-hand side of (5.17) becomes smaller or equal to zero for the first time.

In order to show the convergence (5.16), we note that $\|Ax_p - y^\delta\|_2 \leq \delta$ implies

$$\|Ax_p - y\|_2 \leq \|Ax_p - y^\delta\|_2 + \|y^\delta - y\|_2 \leq 2\delta \to 0, \quad \delta \to 0,$$

i.e., $Ax_p \to y$, $\delta \to 0$. From this, since $A^\dagger A v = v$ for all $v \in \operatorname{span}\{v_1, \ldots, v_r\}$, we finally can conclude that $x_p \to A^\dagger y$, $\delta \to 0$. □

The spectral cutoff method requires the full solution of the eigenvalue problem for the matrix A^*A, which we will describe in Chapter 7. As an alternative, in the following section we shall describe the Tikhonov regularization, which can be performed without explicitly knowing the singular value decomposition.

5.3 Tikhonov Regularization

Tikhonov regularization as introduced independently by Phillips in 1962 and Tikhonov 1963 is obtained from (5.11) by multiplying $1/\mu_j$ by the damping factor

$$\frac{\mu_j^2}{\alpha + \mu_j^2},$$

where α is some positive *regularization parameter.*

Theorem 5.7 *Let A be an $m \times n$ matrix of rank r with singular system (μ_j, u_j, v_j) and let $\alpha > 0$. Then for each $y \in \mathbb{C}^m$ the linear system*

$$\alpha x_\alpha + A^*Ax_\alpha = A^*y \tag{5.18}$$

is uniquely solvable, and the solution is given by

$$x_\alpha = \sum_{j=1}^{r} \frac{\mu_j}{\alpha + \mu_j^2}\,(y, v_j)\,u_j. \tag{5.19}$$

Proof. For $\alpha > 0$ the matrix $\alpha I + A^*A$ is positive definite and therefore nonsingular. Since

$$\alpha u_j + A^*Au_j = (\alpha + \mu_j^2)u_j,$$

a singular system for the matrix $\alpha I + A^*A$ is given by $(\alpha + \mu_j^2, u_j, u_j)$, $j = 1, \ldots, n$. Now the assertion follows from Theorem 5.5 with the aid of $(A^*y, u_j) = (y, Au_j)$ and using (5.5). □

Corollary 5.8 *Under the assumptions of Theorem 5.7 we have convergence:*

$$\lim_{\alpha \to 0} (\alpha I + A^* A)^{-1} A^* y = A^\dagger y.$$

Proof. This is obvious from (5.13) and (5.19). □

Before we proceed with a discussion on how to choose the regularization parameter α, we give an interpretation of Tikhonov regularization as a penalized least squares method.

Theorem 5.9 *Let A be an $m \times n$ matrix and let $\alpha > 0$. Then for each $y \in \mathbb{C}^m$ there exists a unique $x_\alpha \in \mathbb{C}^n$ such that*

$$\|Ax_\alpha - y\|_2^2 + \alpha \|x_\alpha\|_2^2 = \inf_{x \in \mathbb{C}^n} \{\|Ax - y\|_2^2 + \alpha \|x\|_2^2\}. \tag{5.20}$$

The minimizing vector x_α is given by the unique solution of the linear system (5.18).

Proof. (Compare to the proof of Theorem 3.51.) We first note the relation

$$\begin{aligned} \|Ax - y\|_2^2 + \alpha\|x\|_2^2 &= \|Ax_\alpha - y\|_2^2 + \alpha\|x_\alpha\|_2^2 \\ &\quad + 2\,\mathrm{Re}\,(x - x_\alpha, \alpha x_\alpha + A^* A x_\alpha - A^* y) \\ &\quad + \|Ax - Ax_\alpha\|_2^2 + \alpha\|x - x_\alpha\|_2^2, \end{aligned} \tag{5.21}$$

which is valid for all $x, x_\alpha \in \mathbb{C}^n$. From this it is obvious that the solution x_α of (5.18) satisfies (5.20).

Conversely, let x_α be a solution of (5.20) and assume that

$$\alpha x_\alpha + A^* A x_\alpha \neq A^* y.$$

Then, setting $z := \alpha x_\alpha + A^* A x_\alpha - A^* y$, for $x := x_\alpha - \varepsilon z$ with $\varepsilon \in \mathbb{R}$ from (5.21) we have

$$\|Ax - y\|_2^2 + \alpha\|x\|_2^2 = \|Ax_\alpha - y\|_2^2 + \alpha\|x_\alpha\|_2^2 - 2\varepsilon a + \varepsilon^2 b,$$

where

$$a := \|z\|_2^2 \quad \text{and} \quad b := \|Az\|_2^2 + \alpha\|z\|_2^2$$

are both positive. By choosing $\varepsilon = a/b$ we obtain

$$\|Ax - y\|_2^2 + \alpha\|x\|_2^2 < \|Ax_\alpha - y\|_2^2 + \alpha\|x\|_2^2,$$

which contradicts (5.20). □

The interpretation of Tikhonov regularization through the above Theorem 5.9 indicates that it keeps the residual $\|Ax_\alpha - y\|_2^2$ small and stabilizes by preventing x_α from becoming large through the penalty term $\alpha\|x_\alpha\|_2^2$.

From the proof of Theorem 5.7 we know that the eigenvalues of the Hermitian matrix $\alpha I + A^*A$ are given by $\alpha + \mu_j^2$, $j = 1, \dots, n$. Hence by Theorem 3.27 we have that

$$\operatorname{cond}_2(\alpha I + A^*A) = \frac{\alpha + \mu_1^2}{\alpha + \mu_n^2} \le \frac{2\mu_1^2}{\alpha}, \quad 0 < \alpha \le \mu_1^2. \tag{5.22}$$

Therefore stability of the linear system (5.18) requires the regularization parameter α to be fairly large. On the other hand, in order to keep the system (5.18) reasonably close to the original system $Ax = y$, we expect that α needs to be small. This observation is made more precise through the following considerations on the error occurring in Tikhonov regularization.

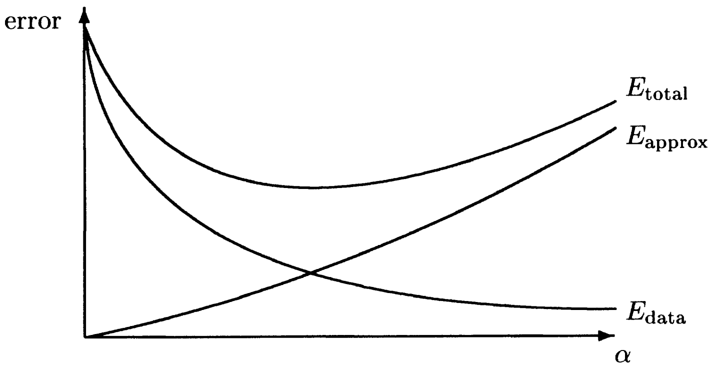


FIGURE 5.1. Total error for Tikhonov regularization

Given an erroneous right-hand side y^δ with error level $\|y^\delta - y\|_2 \le \delta$, the Tikhonov regularization approximates the solution $x = A^\dagger y$ of $Ax = y$ by the solution x_α of the regularized linear system

$$\alpha x_\alpha + A^*Ax_\alpha = A^*y^\delta. \tag{5.23}$$

Then, for the total error, writing

$$x_\alpha - x = (\alpha I + A^*A)^{-1}A^*(y^\delta - y) + (\alpha I + A^*A)^{-1}A^*y - A^\dagger y,$$

by the triangle inequality we have the estimate

$$\|x_\alpha - x\|_2 \le \|(\alpha I + A^*A)^{-1}A^*\|_2\,\delta + \|(\alpha I + A^*A)^{-1}A^*y - A^\dagger y\|_2.$$

This decomposition shows that the total error consists of two parts:

$$E_{\text{total}} \le E_{\text{data}} + E_{\text{approx}}.$$

The first term, with the aid of Theorem 3.31, can be estimated by

$$E_{\text{data}} = \|(\alpha I + A^*A)^{-1}A^*\|_2\,\delta \ge \frac{\mu_r}{\alpha + \mu_r^2}\,\delta.$$

It reflects the influence of the incorrect data and, for fixed δ, becomes large as $\alpha \to 0$, if the smallest positive singular value μ_r is close to zero (see also Problem 5.16). The second term,

$$E_{\text{approx}} = \|(\alpha I + A^* A)^{-1} A^* y - A^\dagger y\|_2,$$

describes the approximation error due to the replacement of $Ax = y$ by the regularized equation (5.23), and by Corollary 5.8, it goes to zero as $\alpha \to 0$. This error behavior is illustrated in Figure 5.1.

On one hand, in view of (5.22) the stability of the system requires a large regularization parameter α to keep E_{data} small, i.e., to keep the influence of the data error $\|y^\delta - y\|_2$ small. On the other hand, keeping E_{approx} small asks for a small parameter α.

Obviously, the choice of the parameter α has to be made through a compromise between accuracy and stability. An efficient strategy to achieve this is again provided by the discrepancy principle. In the following theorem we need to assume that $Ax = y$ is solvable.

Theorem 5.10 *Let A be an $m \times n$ matrix and let $y \in A(\mathbb{C}^n)$, $y^\delta \in \mathbb{C}^m$ satisfy*

$$\|y^\delta - y\|_2 \leq \delta < \|y^\delta\|_2$$

for $\delta > 0$. Then there exists a unique $\alpha = \alpha(\delta) > 0$ such that the unique solution x_α of (5.23) satisfies

$$\|Ax_\alpha - y^\delta\|_2 = \delta. \tag{5.24}$$

This discrecancy principle for Tikhonov regularization is regular in the sense that if the error level δ tends to zero, then

$$x_\alpha \to A^\dagger y, \quad \delta \to 0. \tag{5.25}$$

Proof. We have to show that the function $F : (0, \infty) \to \mathbb{R}$ defined by

$$F(\alpha) := \|Ax_\alpha - y^\delta\|_2^2 - \delta^2$$

has a unique zero. In terms of a singular system, from the representation (5.19) we find that

$$F(\alpha) = \sum_{j=1}^{m} \frac{\alpha^2}{(\alpha + \mu_j^2)^2} |(y^\delta, v_j)|^2 - \delta^2.$$

Therefore, F is continuous and strictly monotonically increasing with the limits $F(\alpha) \to -\delta^2 < 0$, $\alpha \to 0$, and $F(\alpha) \to \|y^\delta\|_2^2 - \delta^2 > 0$, $\alpha \to \infty$. Hence, F has exactly one zero $\alpha = \alpha(\delta)$.

Note that the condition $\|y^\delta - y\|_2 \leq \delta < \|y^\delta\|_2$ implies that $y \neq 0$. Using (5.23), (5.24), and the triangle inequality we can estimate

$$\|y^\delta\|_2 - \delta = \|y^\delta\|_2 - \|Ax_\alpha - y^\delta\|_2 \leq \|Ax_\alpha\|_2$$

and

$$\alpha \|Ax_\alpha\|_2 = \|AA^*(y^\delta - Ax_\alpha)\|_2 \le \|AA^*\|_2 \delta.$$

Combining these two inequalities and using $\|y^\delta\|_2 \ge \|y\|_2 - \delta$ yields

$$\alpha \le \frac{\|AA^*\|_2 \, \delta}{\|y\|_2 - 2\delta}.$$

This implies that $\alpha \to 0$, $\delta \to 0$. Now the convergence (5.25) follows from the representations (5.13) for $A^\dagger y$ and (5.19) for x_α (with y replaced by y^δ) and the fact that $\|y^\delta - y\|_2 \to 0$, $\delta \to 0$. □

In practice, of course, one does not need to determine the regularization parameter satisfying (5.24) exactly. Usually the following strategy will be sufficient: Choose some moderately sized α and then keep decreasing α by a constant factor γ, say $\gamma = 0.5$, until $F(\alpha)$ becomes negative.

In order to illustrate that Tikhonov regularization works, Table 5.4 gives some numerical results for the linear system of Example 5.1 with the erroneous right-hand side generated by using $\ln 2 \approx 0.69315$ and choosing the regularizing parameter $\alpha = 10^{-10}$ (without attempting to use Theorem 5.10).

TABLE 5.4. Regularized solution of the linear system (5.1)

n	α_0	α_1	α_2	α_3	α_4	α_5	α_6
1	0.9315	−0.4767					
2	0.9862	−0.8052	0.3285				
3	0.9987	−0.9546	0.7021	−0.2491			
4	1.0015	−1.0193	1.0154	−0.7605	0.2644		
5	0.9992	−0.9659	0.7236	−0.1458	−0.2838	0.1735	
6	0.9995	−0.9618	0.6564	0.0254	−0.2818	−0.1512	0.2166

Problems

5.1 For the condition number of linear operators show that

$$\operatorname{cond}(AB) \le \operatorname{cond}(A)\operatorname{cond}(B).$$

5.2 Let A be an $n \times n$ matrix and Q be a unitary $n \times n$ matrix. Show that

$$\operatorname{cond}_2(QA) = \operatorname{cond}_2(A)$$

and

$$\operatorname{cond}_2(A^*A) \ge \operatorname{cond}_2(A).$$

5.3 Determine $\operatorname{cond}_2(A)$ for the matrix A of Example 2.1 and discuss its behavior for large n.

5.4 Find the inverse of the matrix

$$A = \begin{pmatrix} 5 & 7 & 3 \\ 7 & 11 & 2 \\ 3 & 2 & 6 \end{pmatrix}$$

and find the condition numbers $\text{cond}_p(A)$ for $p = 1, 2, \infty$.

5.5 Find the inverse of the matrix

$$A = \begin{pmatrix} 10 & 1 & 4 & 0 \\ 1 & 10 & 5 & -1 \\ 4 & 5 & 10 & 7 \\ 0 & -1 & 7 & 9 \end{pmatrix}$$

and find the condition numbers $\text{cond}_p(A)$ for $p = 1, 2, \infty$.

5.6 Calculate $\text{cond}_\infty(A)$ for the matrix

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 10 & 100 \\ 1 & 100 & 10000 \end{pmatrix}.$$

Show that one can improve the condition of a matrix by scaling through calculating $\text{cond}_\infty(DA)$ where D is the diagonal matrix

$$D = \text{diag}(1/3, 1/111, 1/10101).$$

5.7 Let $A = (a_{jk})$ be an $n \times n$ matrix satisfying

$$\sum_{k=1}^{n} |a_{jk}| = 1, \quad j = 1, \ldots, n.$$

Show that

$$\text{cond}_\infty(A) \leq \text{cond}_\infty(DA)$$

for all $n \times n$ diagonal matrices D (see Problem 5.6).

5.8 For a nonsingular matrix A show that

$$\frac{1}{\text{cond}(A)} = \frac{1}{\|A\|} \min\{\|B\| : A + B \text{ is singular}\}.$$

This indicates that if a nonsingular matrix has a large condition number, it is close to a singular matrix.

5.9 Show that for an $m \times n$ matrix A the $n \times n$ matrix $A^* A$ is Hermitian and positive semidefinite.

5.10 Find the singular value decomposition of

$$A = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 0 & -1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}.$$

5.11 Show that $A^\dagger y$ is the least squares solution of $Ax = y$ with minimal norm.

5.12 Show that the pseudo-inverse $A^\dagger$ is uniquely determined by the properties

$$AA^\dagger = (AA^\dagger)^*, \quad A^\dagger A = (A^\dagger A)^*, \quad AA^\dagger A = A, \quad A^\dagger AA^\dagger = A^\dagger.$$

Express the pseudo-inverse in terms of the decomposition (5.8).

5.13 For the pseudo-inverse show that $(A^\dagger)^\dagger = A$ and $(A^\dagger)^* = (A^*)^\dagger$.

5.14 Give an example to show that in general, $(AB)^\dagger \neq B^\dagger A^\dagger$.

5.15 What is the pseudo-inverse of $A : \mathbb{C}^n \to \mathbb{C}^m$ given by $Ax = (x, a)b$ with $a \in \mathbb{C}^m$ and $b \in \mathbb{C}^n$?

5.16 For an $m \times n$ matrix show that

$$\|(\alpha I + A^*A)^{-1}A^*\|_2 \leq \frac{1}{\sqrt{\alpha}}$$

for $\alpha > 0$.

5.17 Give an alternative proof of Theorem 5.9 by using the necessary and sufficient conditions for the minimum of a function of n variables.

5.18 Let X and Y be finite-dimensional pre-Hilbert spaces and let $A : X \to Y$ be a linear operator. Show that there exists a uniquely determined linear operator $A^* : Y \to X$ with the property

$$(Ax, y)_Y = (x, A^*y)_X$$

for all $x \in X$ and $y \in Y$. Use this result to formulate and prove a generalization of Theorem 5.9 for the minimization of

$$\|Ax - y\|_Y^2 + \alpha\|x\|_X^2.$$

5.19 Show that

$$(x, y) := \sum_{j=0}^{n} x_j \bar{y}_j + \sum_{j=1}^{n} (x_j - x_{j-1})(\bar{y}_j - \bar{y}_{j-1})$$

defines a scalar product on $\mathbb{C}^n$. Discuss its use in Tikhonov regularization as indicated in Problem 5.18, where in addition to large components of the solution vector oscillations between consecutive components are also penalized.

5.20 Show that $A : C[0, 1] \to C[0, 1]$ defined by

$$(Af)(x) := \int_0^x f(y)\,dy, \quad x \in [0, 1],$$

is a bounded linear operator that does not have a bounded inverse; i.e., show that differentiation is an ill-posed problem.

6

Iterative Methods for Nonlinear Systems

In this chapter we will study the solution of systems of nonlinear equations. As opposed to linear equations, no explicit solution techniques are, in general, available for nonlinear equations, and hence their solution completely relies on iterative methods. In the first section we shall begin with the application of the Banach fixed point theorem for systems of nonlinear equations with one or several variables. Given the fact that iterative techniques have a long history in mathematics, the significance of Banach's fixed point theorem originates from its unified approach, covering a wide variety of different successive approximation methods.

In the second section, we will continue with the study of Newton's iteration method for finding zeros of functions of one or several variables. This iteration scheme is attributed to Newton, since in 1669 he developed a solution method for cubic equations by linearization that may be viewed as a precursor of what is now known as Newton iteration. He also used this method for approximately solving Kepler's equations for planetary motion.

In the concluding two sections of this chapter we will consider the application of Newton's method for finding zeros of polynomials and its modification into the more recently developed Levenberg–Marquardt scheme for solving the least squares problem.

Given the vast number of iterative methods available for nonlinear equations, we will confine our presentation to describing the fundamental ideas and will not aim at a complete treatment of the subject.

6.1 Successive Approximations

In this section, we will consider systems of n nonlinear equations for n unknowns of the form

$$f(x) = x,$$

where $x = (x_1, \dots, x_n)^T$ and $f(x) = (f_1(x_1, \dots, x_n), \dots, f_n(x_1, \dots, x_n))^T$. We begin by studying the case of a single nonlinear equation with one unknown. Obviously, in one dimension, solving $f(x) = x$ geometrically corresponds to determining the intersection of the graph of the function f with the straight line described by the function $x \mapsto x$.

Theorem 6.1 *Let $D \subset \mathbb{R}$ be a closed interval and let $f : D \to D$ be a continuously differentiable function with the property*

$$q := \sup_{x \in D} |f'(x)| < 1.$$

Then the equation $f(x) = x$ has a unique solution $x \in D$, and the successive approximations

$$x_{\nu+1} := f(x_\nu), \quad \nu = 0, 1, 2, \dots,$$

with arbitrary $x_0 \in D$ converge to this solution. We have the a priori error estimate

$$|x_\nu - x| \le \frac{q^\nu}{1-q} |x_1 - x_0|$$

and the a posteriori error estimate

$$|x_\nu - x| \le \frac{q}{1-q} |x_\nu - x_{\nu-1}|$$

for all $\nu \in \mathbb{N}$.

Proof. Equipped with the norm $\|\cdot\| = |\cdot|$ the space $\mathbb{R}$ is complete. By the mean value theorem, for $x, y \in D$ with $x < y$, we have that

$$f(x) - f(y) = f'(\xi)(x - y)$$

for some intermediate point $\xi \in (x, y)$. Hence

$$|f(x) - f(y)| \le \sup_{\xi \in D} |f'(\xi)| \, |x - y| = q|x - y|,$$

which is also valid for $x, y \in D$ with $x \ge y$. Therefore, f is a contraction, and the assertion follows from the Banach fixed point Theorem 3.46. □

Figure 6.1 illustrates graphically the successive approximations for functions f with positive and negative slope, respectively, of absolute value less than one. Note that the sequence (x_ν) converges to the fixed point

monotonically if f has positive slope and that it converges with values alternating above and below the fixed point if f has negative slope. In both cases the slope of the function f has absolute value less than one in a neighborhood of the fixed point. From drawing a corresponding figure for a function with a slope of absolute value greater than one it can be seen that the corresponding iteration will move away from the fixed point (see Problem 6.2).

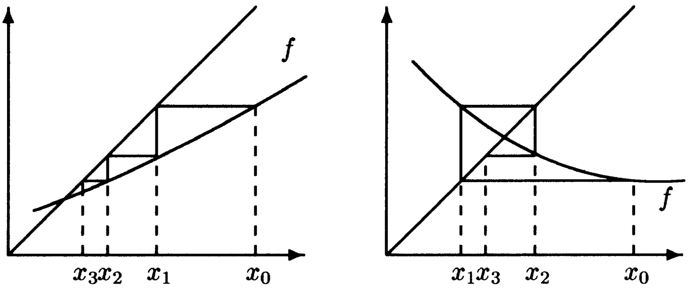


FIGURE 6.1. Fixed point iteration

The following theorem states that for a fixed point x with $|f'(x)| < 1$ we always can find starting points x_0 ensuring convergence of the successive approximations.

Theorem 6.2 *Let x be a fixed point of a continuously differentiable function f such that $|f'(x)| < 1$. Then the method of successive approximations $x_{\nu+1} := f(x_\nu)$ is* locally convergent*; i.e., there exists a neighborhood B of the fixed point x such that the successive approximations converge to x for all $x_0 \in B$.*

Proof. Since f' is continuous and $|f'(x)| < 1$, there exist constants $0 < q < 1$ and $\delta > 0$ such that $|f'(y)| \leq q$ for all $y \in B := [x - \delta, x + \delta]$. Then we have that

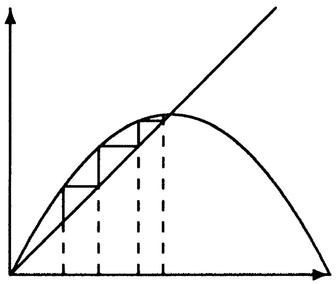
$$|f(y) - x| = |f(y) - f(x)| \leq q|y - x| \leq |y - x| \leq \delta$$

for all $y \in B$; i.e., f maps B into itself and is a contraction $f : B \to B$. Now the statement of the theorem follows from Theorem 6.1. □

Theorem 6.2 expresses the fact that for a fixed point x with $|f'(x)| < 1$ the sequence $x_{\nu+1} := f(x_\nu)$ converges if the starting point x_0 is sufficiently close to x. In practical situations the problem of how to obtain such a good initial guess is unresolved in general. Frequently, however, a good estimate of the fixed point might be known a priori from the underlying application or might be deduced from analytic observations.

The following examples illustrate that in some cases we also have *global convergence*, where the successive approximations converge for each starting point in the domain of definition of the function f.

Example 6.3 In order to describe a division by iteration, for $a > 0$ we consider the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) := 2x - ax^2$. The graph of this function is a parabola with maximum value $1/a$ attained at $1/a$. By solving the quadratic equation $f(x) = x$ it can be seen that f has the fixed points $x = 0$ and $x = 1/a$. Obviously, f maps the open interval $(0, 2/a)$ into $(0, 1/a)$. Since $f'(x) = 2(1 - ax)$, we have $f'(0) = 2$ and $f'(1/a) = 0$. From the the property $x < f(x) < 1/a$, which is valid for $0 < x < 1/a$, it follows that the sequence $x_{\nu+1} := 2x_\nu - ax_\nu^2$ is monotonicly increasing and bounded. Hence, the successive approximations converge to the fixed point $x = 1/a$ for arbitrarily chosen $x_0 \in (0, 2/a)$. Figure 6.2 illustrates the convergence. The numerical results are for $a = 2$ and two different starting points, $x_0 = 0.3$ and $x_0 = 0.4$. □



ν	x_ν	x_ν
0	0.30000000	0.40000000
1	0.42000000	0.48000000
2	0.48720000	0.49920000
3	0.49967232	0.49999872

FIGURE 6.2. Division by iteration

Example 6.4 For computing the square root of a positive real number a by an iterative method we consider the function $f : (0, \infty) \to (0, \infty)$ given by

$$f(x) := \frac{1}{2}\left(x + \frac{a}{x}\right).$$

By solving the quadratic equation $f(x) = x$ it can be seen that f has the fixed point $x = \sqrt{a}$. By the arithmetic geometric mean inequality we have that $f(x) > \sqrt{a}$ for $x > 0$; i.e., f maps the open interval $(0, \infty)$ into $[\sqrt{a}, \infty)$, and therefore it maps the closed interval $[\sqrt{a}, \infty)$ into itself. From

$$f'(x) = \frac{1}{2}\left(1 - \frac{a}{x^2}\right)$$

it follows that

$$q := \sup_{\sqrt{a} \le x < \infty} |f'(x)| = \frac{1}{2}.$$

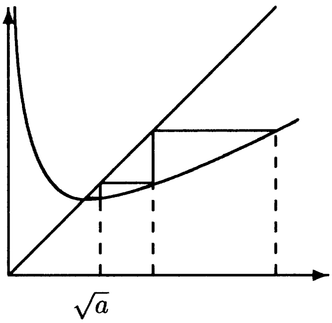
Hence $f : [\sqrt{a}, \infty) \to [\sqrt{a}, \infty)$ is a contraction. Therefore, by Theorem 6.1 the successive approximations

$$x_{\nu+1} := \frac{1}{2}\left(x_\nu + \frac{a}{x_\nu}\right), \quad \nu = 0, 1, \dots,$$

converge to the square root $\sqrt{a}$ for each $x_0 > 0$, and we have the a posteriori error estimate

$$|\sqrt{a} - x_\nu| \leq |x_\nu - x_{\nu-1}|.$$

Figure 6.3 illustrates the convergence. The numerical results again are for $a = 2$. □



ν	x_ν
0	5.00000000
1	2.70000000
2	1.72037037
3	1.44145537
4	1.41447098
5	1.41421359
6	1.41421356

FIGURE 6.3. Square root by iteration

In both of Examples 6.3 and 6.4 the numerical values exhibit a very rapid convergence. This is due to the fact that because of $f'(x) = 0$ at the fixed point, the contraction number is very small. We shall elaborate on this observation later when we consider Newton's method.

TABLE 6.1. Iterations for Example 6.5

ν	x_ν		ν	x_ν
0	1.00000000		7	0.72210243
1	0.54030231		·	·
2	0.85755322		·	·
3	0.65428979		45	0.73908513
4	0.79348036		46	0.73908514
5	0.70136877		47	0.73908513
6	0.76395968		48	0.73908513

Example 6.5 Consider the function $f : [0,1] \to [0,1]$ given by

$$f(x) := \cos x.$$

Here we have

$$q = \sup_{0 \leq x \leq 1} |f'(x)| = \sin 1 < 1,$$

and Theorem 6.1 implies that the successive approximations $x_{\nu+1} := \cos x_\nu$ converge to the unique solution x of $\cos x = x$ for each $x_0 \in [0, 1]$. Table 6.1 illustrates the convergence, which is notably slower than in the two previous examples. □

By the following example we illustrate how to obtain a fixed point of a function with derivative greater than one by working with the inverse function.

Example 6.6 The function $h : (0, 1) \to (-\infty, \infty)$ given by $h(x) := x + \ln x$ is strictly monotonically increasing with limits $\lim_{x\to 0} h(x) = -\infty$ and $\lim_{x\to\infty} h(x) = \infty$. Therefore, the function $f(x) := -\ln x$ has a unique fixed point x. Since this fixed point must satisfy $0 < x < 1$, the derivative

$$|f'(x)| = \frac{1}{x} > 1$$

implies that f is not contracting in a neighborhood of the fixed point. However, we can still design a convergent scheme because $x = -\ln x$ is equivalent to $e^{-x} = x$. We consider the inverse function

$$g(x) := e^{-x}$$

of f, which has derivative $|g'(x)| = e^{-x} < 1$ at the fixed point, so that we can apply Theorem 6.2. Obviously, for each $0 < a < 1/e$ the exponential function g maps the interval $[a, 1]$ into itself. Since

$$q = \sup_{a \le x \le 1} |g'(x)| = e^{-a} < 1,$$

by Theorem 6.1 it follows that for arbitrary $x_0 > 0$ the successive approximations $x_{\nu+1} = e^{-x_\nu}$ converge to the unique solution of $x = e^{-x}$. □

Now we will extend Theorem 6.1 to systems of nonlinear equations. A subset D of a linear space X is called *convex* if

$$\lambda x + (1 - \lambda) y \in D$$

for all $x, y \in D$ and all $\lambda \in (0, 1)$, i.e., if the straight line connecting x and y is contained in D.

Theorem 6.7 *Let $D \subset \mathbb{R}^n$ be open and convex and let $f : D \to \mathbb{R}^n$ be a mapping*

$$f(x) = (f_1(x_1, \ldots, x_n), \ldots, f_n(x_1, \ldots, x_n))^T,$$

where the $f_j : D \to \mathbb{R}$, $j = 1, \ldots, n$, are continuously differentiable functions. By

$$f'(x) = \left(\frac{\partial f_j}{\partial x_k}(x)\right)_{j,k=1,\ldots,n}$$

we denote the Jacobian matrix *of f. Then we have the* mean value theorem

$$\|f(x) - f(y)\| \le \max_{0 \le \lambda \le 1} \|f'[\lambda x + (1-\lambda)y]\| \, \|x - y\|$$

for all $x, y \in D$ (and all norms $\|\cdot\|$ on $\mathbb{R}^n$).

Proof. Let $g : [0,1] \to \mathbb{R}^n$ be continuous. We will show that

$$\left\| \int_0^1 g(\lambda)\, d\lambda \right\| \le \int_0^1 \|g(\lambda)\|\, d\lambda, \tag{6.1}$$

where the integral on the left-hand side has to be understood as the vector of the integrals over the components of g. The function $\lambda \mapsto \|g(\lambda)\|$ is continuous, since the norm is a continuous function. Therefore, the integral on the right-hand side of (6.1) is well-defined. Consider the equidistant subdivision $\lambda_i = i/m$, $i = 0, 1, \ldots, m$, for $m \in \mathbb{N}$. Then we have the converging Riemann sums

$$\sum_{i=1}^m \|g(\lambda_i)\| \, (\lambda_i - \lambda_{i-1}) \to \int_0^1 \|g(\lambda)\|\, d\lambda, \quad m \to \infty,$$

and

$$\sum_{i=1}^m g(\lambda_i) \, (\lambda_i - \lambda_{i-1}) \to \int_0^1 g(\lambda)\, d\lambda, \quad m \to \infty.$$

From the second limit, by the continuity of the norm we conclude that

$$\left\| \sum_{i=1}^m g(\lambda_i) \, (\lambda_i - \lambda_{i-1}) \right\| \to \left\| \int_0^1 g(\lambda)\, d\lambda \right\|, \quad m \to \infty.$$

Now (6.1) follows by passing to the limit $m \to \infty$ in the inequality

$$\left\| \sum_{i=1}^m g(\lambda_i) \, (\lambda_i - \lambda_{i-1}) \right\| \le \sum_{i=1}^m \|g(\lambda_i)\| \, (\lambda_i - \lambda_{i-1}),$$

which is a consequence of the triangle inequality.

Since D is convex, for all $x, y \in D$ we have that

$$f_j(x) - f_j(y) = \int_0^1 \frac{d}{d\lambda} f_j[\lambda x + (1-\lambda)y]\, d\lambda, \quad j = 1, \ldots, n.$$

By the chain rule we compute

$$\frac{d}{d\lambda} f_j[\lambda x + (1-\lambda)y] = \sum_{k=1}^n \frac{\partial f_j}{\partial x_k} [\lambda x + (1-\lambda)y] \, (x_k - y_k),$$

and therefore

$$f_j(x) - f_j(y) = \int_0^1 \sum_{k=1}^n \frac{\partial f_j}{\partial x_k} [\lambda x + (1-\lambda)y] (x_k - y_k) \, d\lambda;$$

i.e., in vector form,

$$f(x) - f(y) = \int_0^1 f'[\lambda x + (1-\lambda)y] (x - y) \, d\lambda.$$

From this, with the aid of (6.1) and the continuity of $\lambda \mapsto f'[\lambda x + (1-\lambda)y]$, we obtain

$$\begin{aligned} \|f(x) - f(y)\| &\leq \int_0^1 \|f'[\lambda x + (1-\lambda)y]\| \, \|x - y\| \, d\lambda \\ &\leq \max_{0 \leq \lambda \leq 1} \|f'[\lambda x + (1-\lambda)y]\| \, \|x - y\|, \end{aligned}$$

which ends the proof. □

Theorem 6.8 *Let $D \subset \mathbb{R}^n$ be closed and convex (with a nonempty interior) and let $f : D \to D$ be a continuous mapping. Assume further that f is continuously differentiable in the interior of D and that its Jacobian can be continuously extended to all of D such that*

$$\sup_{x \in D} \|f'(x)\| < 1$$

in some norm $\|\cdot\|$ on $\mathbb{R}^n$. Then the equation $f(x) = x$ has a unique solution $x \in D$, and the successive approximations

$$x_{\nu+1} := f(x_\nu), \quad \nu = 0, 1, 2, \ldots,$$

converge for each $x_0 \in D$ to this fixed point. We have the a priori error estimate

$$\|x_\nu - x\| \leq \frac{q^\nu}{1-q} \|x_1 - x_0\|$$

and the a posteriori error estimate

$$\|x_\nu - x\| \leq \frac{q}{1-q} \|x_\nu - x_{\nu-1}\|$$

for all $\nu \in \mathbb{N}$.

Proof. By the mean value Theorem 6.7 the mapping $f : D \to D$ is a contraction. □

By Theorem 3.26 we have that each of the conditions

$$\sup_{x\in D} \max_{j=1,\ldots,n} \sum_{k=1}^{n} \left| \frac{\partial f_j}{\partial x_k}(x) \right| < 1,$$

$$\sup_{x\in D} \max_{k=1,\ldots,n} \sum_{j=1}^{n} \left| \frac{\partial f_j}{\partial x_k}(x) \right| < 1,$$

$$\sup_{x\in D} \left[\sum_{j,k=1}^{n} \left| \frac{\partial f_j}{\partial x_k}(x) \right|^2, \right]^{1/2} < 1$$

ensures convergence of the successive approximations in Theorem 6.8.

The following local convergence theorem can be proven analogously to Theorem 6.2.

Theorem 6.9 *Let x be a fixed point of a continuously differentiable function f such that $\|f'(x)\| < 1$ in some norm $\|\cdot\|$ on $\mathbb{R}^n$. Then the method of successive approximations $x_{\nu+1} := f(x_\nu)$ is locally convergent; i.e., there exists a neighborhood B of the fixed point x such that the successive approximations converge to x for all starting elements $x_0 \in B$.*

Example 6.10 For the system

$$x_1 = 0.5\cos x_1 - 0.5\sin x_2$$

$$x_2 = 0.5\sin x_1 + 0.5\cos x_2$$

we have

$$f'(x) = \begin{pmatrix} -0.5\sin x_1 & -0.5\cos x_2 \\ 0.5\cos x_1 & -0.5\sin x_2 \end{pmatrix},$$

and therefore $\|f'(x)\|_2 \leq \sqrt{0.5}$ for all $x \in \mathbb{R}^2$. Hence Theorem 6.8 is applicable. □

The reader will not be surprised to learn that for speeding up convergence of the successive approximations, concepts developed for linear equations like relaxation methods or multigrid methods can also be successfully employed in the nonlinear case. However, since we discussed these methods in some detail in Sections 4.2 and 4.3 for linear equations, we shall refrain from repeating the analysis for nonlinear equations.

6.2 Newton's Method

We now want to determine zeros of a function of n variables; i.e., we want to solve equations of the form

$$f(x) = 0,$$

where $f : D \to \mathbb{R}^n$ is a continuously differentiable function defined on some open subset $D \subset \mathbb{R}^n$.

We begin by considering a function of one variable. Let x_0 be an approximation to a zero of the function f. In a neighborhood of x_0, by Taylor's formula we have that

$$f(x) \approx f(x_0) + f'(x_0)\,(x - x_0) =: g(x). \tag{6.2}$$

Therefore, we may consider the zero of the affine linear function g as a new approximation to the zero of f and denote it by x_1. From the linear equation

$$f(x_0) + f'(x_0)\,(x_1 - x_0) = 0 \tag{6.3}$$

we immediately obtain

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}\,.$$

Geometrically, the affine linear function g describes the tangent line to the graph of the function f at the point x_0.

This consideration can be extended to the case of more than one variable. Given an approximation x_0 to a zero of f, by Taylor's formula we still have the approximation (6.2), where now, as in the previous section,

$$f'(x) = \left(\frac{\partial f_j}{\partial x_k}\,(x)\right)_{j,k=1,\dots,n}$$

denotes the Jacobian matrix of f. Again we obtain a new approximation x_1 for the solution of $f(x) = 0$ by solving the linearized equation (6.3), i.e., by

$$x_1 = x_0 - [f'(x_0)]^{-1} f(x_0).$$

Geometrically, the function g of (6.2) corresponds to the hyperplane tangent to f at the point x_0.

Iterating this procedure leads to Newton's method, as described in the following definition. In the case of one variable, the geometric situation is shown in Figure 6.4.

Definition 6.11 *Let $D \subset \mathbb{R}^n$ be open and let $f : D \to \mathbb{R}^n$ be a continuously differentiable function such that the Jacobian matrix $f'(x)$ is nonsingular for all $x \in D$. Then* Newton's method *for the solution of the equation*

$$f(x) = 0$$

is given by the iteration scheme

$$x_{\nu+1} := x_\nu - [f'(x_\nu)]^{-1} f(x_\nu), \quad \nu = 0, 1, \dots,$$

starting with some $x_0 \in D$.

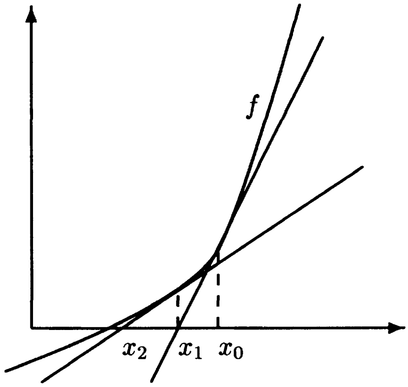


FIGURE 6.4. Newton's method

We explicitly note that $x_{\nu+1}$ is obtained by solving the system of linear equations

$$f'(x_\nu)\,(x_\nu - x_{\nu+1}) = f(x_\nu)$$

for $x_\nu - x_{\nu+1}$; i.e., no matrix inversion is required.

Example 6.12 For the function

$$f(x) := a - \frac{1}{x}$$

where $a > 0$, the Newton iteration is given by

$$x_{\nu+1} := 2x_\nu - ax_\nu^2.$$

By Example 6.3 we have convergence for all $x_0 \in (0, 2/a)$. □

Example 6.13 For the function

$$f(x) =: x^2 - a$$

where $a > 0$, the Newton iteration is given by

$$x_{\nu+1} := \frac{1}{2}\left(x_\nu + \frac{a}{x_\nu}\right).$$

By Example 6.4 we have convergence for all $x_0 \in (0, \infty)$. □

Of course, we cannot expect that Newton method's will always converge. However, by the following analysis we can assure local convergence.

Theorem 6.14 *Let $D \subset \mathbb{R}^n$ be open and convex and let $f : D \to \mathbb{R}^n$ be continuously differentiable. Assume that for some norm $\|\cdot\|$ on $\mathbb{R}^n$ and some $x_0 \in D$ the following conditions hold:*

(*a*) *f satisfies*

$$\|f'(x) - f'(y)\| \le \gamma \|x - y\|$$

for all $x, y \in D$ *and some constant* $\gamma > 0$.

(*b*) *The Jacobian matrix* $f'(x)$ *is nonsingular for all* $x \in D$, *and there exists a constant* $\beta > 0$ *such that*

$$\|[f'(x)]^{-1}\| \le \beta, \quad x \in D.$$

(*c*) *For the constants*

$$\alpha := \|[f'(x_0)]^{-1} f(x_0)\| \quad \text{and} \quad q := \alpha\beta\gamma$$

the inequality

$$q < \frac{1}{2}$$

is satisfied.

(*d*) *For* $r := 2\alpha$ *the closed ball* $B[x_0, r] := \{x : \|x - x_0\| \le r\}$ *is contained in* D.

Then f *has a unique zero* x^* *in* $B[x_0, r]$. *Starting with* x_0 *the Newton iteration*

$$x_{\nu+1} := x_\nu - [f'(x_\nu)]^{-1} f(x_\nu), \quad \nu = 0, 1, \ldots, \tag{6.4}$$

is well-defined. The sequence (x_ν) *converges to the zero* x^* *of* f, *and we have the error estimate*

$$\|x_\nu - x^*\| \le 2\alpha q^{2^\nu - 1}, \quad \nu = 0, 1, \ldots.$$

Proof. **1.** Let $x, y, z \in D$. From the proof of Theorem 6.7 we know that

$$f(y) - f(x) = \int_0^1 f'[\lambda x + (1 - \lambda)y]\,(y - x)\,d\lambda.$$

Hence

$$f(y) - f(x) - f'(z)\,(y - x) = \int_0^1 \{f'[\lambda x + (1 - \lambda)y] - f'(z)\}\,(y - x)\,d\lambda,$$

and estimating with the aid of (6.1) and condition (a) we find that

$$\begin{aligned} &\|f(y) - f(x) - f'(z)\,(y - x)\| \\ &\le \gamma\|y - x\| \int_0^1 \|\lambda(x - z) + (1 - \lambda)(y - z)\|\,d\lambda \\ &\le \frac{\gamma}{2}\,\|y - x\|\,\{\|x - z\| + \|y - z\|\}. \end{aligned}$$

Choosing $z = x$ shows that

$$\|f(y) - f(x) - f'(x)(y - x)\| \leq \frac{\gamma}{2}\|y - x\|^2 \tag{6.5}$$

for all $x, y \in D$, and choosing $z = x_0$ yields

$$\|f(y) - f(x) - f'(x_0)(y - x)\| \leq r\gamma\|y - x\| \tag{6.6}$$

for all $x, y \in B[x_0, r]$.

2. We proceed by proving through induction that

$$\|x_\nu - x_0\| \leq r \quad \text{and} \quad \|x_\nu - x_{\nu-1}\| \leq \alpha q^{2^{\nu-1}-1}, \quad \nu = 1, 2, \ldots. \tag{6.7}$$

This is valid for $\nu = 1$, since

$$\|x_1 - x_0\| = \|[f'(x_0)]^{-1} f(x_0)\| = \alpha = \frac{r}{2} < r$$

as a consequence of conditions (c) and (d). Assume that the inequalities (6.7) are proven up to some $\nu \geq 1$. Then by condition (b) and since $x_\nu \in B[x_0, r] \subset D$, the element $x_{\nu+1}$ is well-defined. With the aid of condition (b), the definition (6.4) applied to x_ν, the estimate (6.5), the induction assumption, and the definition of q we can estimate

$$\begin{aligned}
\|x_{\nu+1} - x_\nu\| &= \|[f'(x_\nu)]^{-1} f(x_\nu)\| \leq \beta\|f(x_\nu)\| \\
&= \beta\|f(x_\nu) - f(x_{\nu-1}) - f'(x_{\nu-1})(x_\nu - x_{\nu-1})\| \\
&\leq \frac{\beta\gamma}{2}\|x_\nu - x_{\nu-1}\|^2 \leq \frac{\beta\gamma}{2}\left[\alpha q^{2^{\nu-1}-1}\right]^2 = \frac{\alpha}{2}\, q^{2^\nu-1} < \alpha q^{2^\nu-1}.
\end{aligned}$$

From this, with the help of the triangle inequality, the induction assumption, and condition (c), we obtain that

$$\begin{aligned}
\|x_{\nu+1} - x_0\| &\leq \|x_{\nu+1} - x_\nu\| + \cdots + \|x_1 - x_0\| \\
&\leq \alpha\left(1 + q + q^3 + q^7 + \cdots + q^{2^\nu-1}\right) \leq \frac{\alpha}{1-q} \leq 2\alpha = r;
\end{aligned}$$

i.e., the inequalities (6.7) also hold for $\nu + 1$.

3. For $\mu > 0$, using $q < 1/2$, we now can estimate

$$\begin{aligned}
\|x_\nu - x_{\nu+\mu}\| &\leq \|x_\nu - x_{\nu+1}\| + \cdots + \|x_{\nu+\mu-1} - x_{\nu+\mu}\| \\
&\leq \alpha\left(q^{2^\nu-1} + q^{2^{\nu+1}-1} + \cdots + q^{2^{\nu+\mu-1}-1}\right) \\
&= \alpha q^{2^\nu-1}\left(1 + q^{2^\nu} + \cdots + \left[q^{2^\nu}\right]^{2^{\mu-1}-1}\right) \leq 2\alpha q^{2^\nu-1}.
\end{aligned} \tag{6.8}$$

From this we observe that (x_ν) is a Cauchy sequence, since $q < 1/2$ and by Theorem 3.39 the limit

$$x^* = \lim_{\nu\to\infty} x_\nu$$

exists. Passing to the limit $\nu \to \infty$ in (6.7) we obtain $\|x^* - x_0\| \le r$, i.e., $x^* \in B[x_0, r]$, and passing to the limit $\mu \to \infty$ in (6.8) the error estimate of the theorem follows.

4. We now show that the limit x^* is a zero of the function f. With the aid of (6.4) and condition (a) we can estimate

$$\begin{aligned}\|f(x_\nu)\| &= \|f'(x_\nu)\,(x_{\nu+1} - x_\nu)\|\\ &\le \|f'(x_\nu) - f'(x_0) + f'(x_0)\|\,\|x_{\nu+1} - x_\nu\|\\ &\le [\gamma\|x_\nu - x_0\| + \|f'(x_0)\|]\,\|x_{\nu+1} - x_\nu\| \to 0, \quad \nu\to\infty.\end{aligned}$$

Hence $f(x_\nu) \to 0$, $\nu \to \infty$, and the continuity of f implies that indeed $f(x^*) = 0$.

5. We conclude the proof by showing that x^* is the only zero of f in the ball $B[x_0, r]$. For this we consider the function $g : B[x_0, r] \to \mathbb{R}^n$ defined by

$$g(x) := x - [f'(x_0)]^{-1} f(x).$$

From conditions (b) and (c) and the inequality (6.6), by writing

$$g(x) - g(y) = [f'(x_0)]^{-1}\{f(y) - f(x) - f'(x_0)(y - x)\}$$

we deduce that

$$\|g(x) - g(y)\| \le \beta\gamma r\|y - x\| \le 2q\|y - x\|$$

for all $x, y \in B[x_0, r]$; i.e., g is a contraction. Therefore, by Theorem 3.44 the function g has at most one fixed point in $B[x_0, r]$. Now uniqueness of the zero of f in $B[x_0, r]$ follows from the equivalence of the equations $g(x) = x$ and $f(x) = 0$. □

Our main application of Theorem 6.14 consists in deriving the following local convergence result for Newton's method.

Corollary 6.15 *Let $D \subset \mathbb{R}^n$ be open and let $f : D \to \mathbb{R}^n$ be twice continuously differentiable, and assume that x^* is a zero of f such that the Jacobian $f'(x^*)$ is nonsingular. Then Newton's method is locally convergent; i.e., there exists a neighborhood B of the zero x^* such that the Newton iterations converge to x^* for all $x_0 \in B$.*

Proof. Since f is twice continuously differentiable, by the mean value Theorem 6.7 applied to the components of f' there exists $\gamma > 0$ such that

$$\|f'(x) - f'(y)\| \le \gamma\|x - y\|$$

for all x, y in some closed ball $B[x^*, \rho]$ centered at x^*. We write

$$f'(x) = f'(x^*)\{I + [f'(x^*)]^{-1}[f'(x) - f'(x^*)]\}$$

and deduce from the above estimate and Theorem 3.48 that the radius ρ of $B[x^*, \rho]$ can be chosen such that $f'(x)$ is nonsingular on $B[x^*, \rho]$ and $\|[f'(x^*)]^{-1}\| \leq \beta$ for all $x \in B[x^*, \rho]$ and some constant $\beta > 0$.

Since f is continuous, $f(x^*) = 0$ implies that there exists $\delta < \rho/2$ such that

$$\|f(x_0)\| < \min\left\{\frac{\rho}{4\beta}, \frac{1}{2\beta^2\gamma}\right\}$$

for all $\|x_0 - x^*\| < \delta$. Then, after setting $\alpha := \|[f'(x_0)]^{-1}f(x_0)\|$ we have the inequalities

$$\alpha\beta\gamma \leq \|f(x_0)\|\beta^2\gamma < \frac{1}{2}$$

and

$$2\alpha \leq 2\beta\|f(x_0)\| < \frac{\rho}{2}.$$

Hence for the open and convex ball $B(x^*, \rho)$ and for each x_0 with $\|x_0 - x^*\| < \delta$ the assumptions of Theorem 6.14 are satisfied. □

Corollary 6.16 *Let $f : (a, b) \to \mathbb{R}$ be twice continuously differentiable and assume that x^* is a simple zero of f. Then Newton's method is locally convergent.*

Proof. For simple zeros we have $f'(x^*) \neq 0$. □

Example 6.17 For the function $f(x) := x - \cos x$ the Newton iteration reads

$$x_{\nu+1} := x_\nu - \frac{x_\nu - \cos x_\nu}{1 + \sin x_\nu}$$

and leads to the numerical values of Table 6.2. □

TABLE 6.2. Newton iterations for Example 6.17

ν	x_ν
0	1.00000000
1	0.75036387
2	0.73911289
3	0.73908513
4	0.73908513

Example 6.18 For the function $f(x) := x - e^{-x}$ the Newton iteration reads

$$x_{\nu+1} := x_\nu - \frac{x_\nu - e^{-x_\nu}}{1 + e^{-x_\nu}}$$

and leads to the numerical values of Table 6.3. □

TABLE 6.3. Newton iterations for Example 6.18

ν	x_ν
0	1.00000000
1	0.53788284
2	0.56698699
3	0.56714329
4	0.56714329

In both examples we observe that the speed of convergence is considerably improved as compared with the simple successive approximations of Examples 6.5 and 6.6. For a general description of this more rapid convergence of Newton's method we need the following definition.

Definition 6.19 *A convergent sequence (x_ν) from a normed space with limit x is said to be* convergent of order $p \geq 1$ *if there exists a constant $C > 0$ such that*

$$\|x_{\nu+1} - x\| \leq C\|x_\nu - x\|^p, \quad \nu = 1, 2, \ldots.$$

Convergence of order one or two is also called *linear* or *quadratic convergence*, respectively. We note that the convergence in Banach's fixed point Theorem 3.45 is, in general, linear.

Theorem 6.20 *Under the assumptions of Theorem 6.14 Newton's method converges quadratically.*

Proof. Using condition (b) of Theorem 6.14 and the inequality (6.5) we can estimate

$$\begin{aligned}\|x^* - x_{\nu+1}\| &= \|x^* - x_\nu + [f'(x_\nu)]^{-1} f(x_\nu)\| \\ &\leq \|[f'(x_\nu)]^{-1}\| \, \|f(x^*) - f(x_\nu) - f'(x_\nu)(x^* - x_\nu)\| \\ &\leq \frac{\beta\gamma}{2} \|x^* - x_\nu\|^2,\end{aligned}$$

since $f(x^*) = 0$. □

Roughly speaking, the quadratic convergence of Newton's method means that the number of correct digits in the numerical approximation is doubled in each iteration step, as observed in Examples 6.3, 6.4, 6.17, and 6.18. Although by this property Newton's method is very attractive, it has to be observed that one step of the Newton iteration for nonlinear systems can be very costly both through the need for evaluating the entries of the Jacobian $f'(x_\nu)$ and through the cost of solving the linear system to arrive at the new iteration $x_{\nu+1}$. Therefore, a great variety of modifications of Newton's method have been developed that mitigate, in particular, the first difficulty. These *modified Newton methods*, in general, are of the form

$$x_{\nu+1} := x_\nu - A_\nu f(x_\nu), \quad \nu = 0, 1, \dots;$$

i.e., the inverse $[f'(x_\nu)]^{-1}$ of the Jacobian is replaced by some approximating matrix A_ν. Here we will only briefly mention two classical and simple possibilities for avoiding the evaluation of the Jacobian at each iteration step.

In the *simplified,* or *frozen, Newton method,* for all steps the matrix A_ν is kept the same and chosen as the inverse of the Jacobian for the starting point; i.e., the iteration scheme is

$$x_{\nu+1} := x_\nu - [f'(x_0)]^{-1} f(x_\nu), \quad \nu = 0, 1, \dots .$$

Geometrically, in the one-dimensional case this means that the tangent line of f at x_ν is replaced by the parallel to the tangent line of f at x_0 passing through $(x_\nu, f(x_\nu))$.

Theorem 6.21 *Under the assumptions of Theorem 6.14 the simplified Newton method converges linearly to the unique zero of f in $B[x_0, r]$.*

Proof. Recall that the function

$$g(x) := x - [f'(x_0)]^{-1} f(x)$$

defined in the proof of Theorem 6.14 is a contraction. We show that g maps $B[x_0, r]$ into itself. For this we write

$$x_0 - g(x) = [f'(x_0)]^{-1}\{f(x) - f(x_0) - f'(x_0)(x - x_0) + f(x_0)\}.$$

Then estimating with the help of conditions (b), (c) and (d) and the inequality (6.5) we obtain

$$\|g(x) - x_0\| \leq \frac{\beta\gamma}{2} \|x - x_0\|^2 + \alpha \leq 2\alpha^2\beta\gamma + \alpha = (2q + 1)\alpha < 2\alpha = r$$

for all x with $\|x - x_0\| \leq r$. Now the statement of the theorem follows from the Banach fixed point Theorem 3.46. □

In the *secant method* for a function of one variable the derivative $f'(x_\nu)$ is approximated by the difference quotient and the corresponding iterative scheme is given by

$$x_{\nu+1} := x_\nu - \frac{x_\nu - x_{\nu-1}}{f(x_\nu) - f(x_{\nu-1})}\, f(x_\nu), \quad \nu = 0, 1, \dots. \tag{6.9}$$

Geometrically, this means that the tangent line at x_ν is replaced by the secant line through the two points x_ν and $x_{\nu-1}$. Obviously, this method needs two initial elements x_0 and x_1. Generalizations to functions in $\mathbb{R}^n$ are possible (see [47]).

In general, for the simplified Newton method and for the secant method we can expect only linear convergence. The idea underlying the more sophisticated modified Newton methods is to choose the approximating matrices A_ν in a manner leading to an improvement over linear convergence without requiring the computational costs of the full Newton method. In the so called *rank one methods* suggested by Broyden in 1965, in each iteration step the matrix A_ν is updated from the previous matrix $A_{\nu-1}$ by adding only a matrix of rank one such that the resulting iteration scheme is *superlinearly* convergent. Roughly speaking, the latter means that for the sequence $x_\nu \to x$, $\nu \to \infty$, we have that

$$\|x_{\nu+1} - x\| \le C_\nu \|x_\nu - x\|, \quad \nu = 1, 2, \dots,$$

such that $C_\nu \to 0$, $\nu \to \infty$. For details we refer to the literature (see [20, 47]).

6.3 Zeros of Polynomials

In this section we shall apply Newton's method to the computation of the zeros of polynomials. Finding the zeros of polynomials is a classical problem in mathematics and numerical analysis despite the fact that it very seldom occurs in applications. We first observe that Newton's method also works for a complex function of a complex variable, allowing the computation of complex zeros.

Consider the polynomial

$$p(x) = a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_{n-1} x + a_n$$

with real or complex coefficients $a_0, a_1, \dots, a_n$. For the application of Newton's method, in each iteration step we need to compute the values of p and p' at the point x_ν. This can be effectively done by the *Horner scheme.* This is based on writing the polynomial in the form of nested multiplications

$$p(z) = (\cdots((a_0 z + a_1)z + a_2)z + \cdots + a_{n-1})z + a_n,$$

which suggests the recursion

$$b_m = b_{m-1}z + a_m, \quad m = 1, \dots, n, \tag{6.10}$$

starting with $b_0 = a_0$. Performing these n multiplications and additions, we arrive at the value of the polynomial $p(z) = b_n$.

For the polynomial

$$p_1(x) := b_0x^{n-1} + b_1x^{n-2} + b_2x^{n-3} + \cdots + b_{n-2}x + b_{n-1},$$

using (6.10) we compute

$$p_1(x)\,(x-z) + b_n = \sum_{m=0}^{n-1} b_m x^{n-1-m}(x-z) + b_n = \sum_{m=0}^{n} a_m x^{n-m} = p(x).$$

This implies that for a zero z the Horner scheme provides the coefficients of the polynomial obtained by dividing p by the linear factor $x - z$. In addition, we have that

$$p'(x) = p_1'(x)\,(x-z) + p_1(x), \tag{6.11}$$

and in particular,

$$p'(z) = p_1(z).$$

Hence, applying the Horner recursion to the polynomial p_1 yields the value of the derivative $p'(z)$. By repeating this process recursively, we can determine all the derivatives of p at the point z, since by induction, from (6.11) we obtain that

$$p^{(k)}(x) = p_1^{(k)}(x)\,(x-z) + kp_1^{(k-1)}(x),$$

whence

$$p^{(k)}(z) = kp_1^{(k-1)}(z)$$

follows for $k = 1, \dots, n$. Therefore, defining recursively polynomials p_k of degree $n - k$ by applying the Horner scheme to the preceding polynomial p_{k-1} leads to

$$p^{(k)}(z) = k!\,p_k(z), \quad k = 1, \dots, n.$$

We can summarize this in the following theorem.

Theorem 6.22 *Let*

$$p(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \cdots + a_{n-1}x + a_n$$

be a polynomial of degree n. For $z \in \mathbb{C}$ the complete Horner scheme

	a_0	a_1	a_2	$\cdot$	$\cdot$	a_{n-1}	a_n
z	b_0	b_1	b_2	$\cdot$	$\cdot$	b_{n-1}	b_n
z	b'_0	b'_1	b'_2	$\cdot$	$\cdot$	b'_{n-1}	
z	b''_0	b''_1	b''_2	$\cdot$	b''_{n-2}		
$\cdot$	$\cdot$	$\cdot$					
z	$b_0^{(n-1)}$	$b_1^{(n-1)}$					
z	$b_0^{(n)}$						

contains the derivatives

$$b_{n-k}^{(k)} = \frac{p^{(k)}(z)}{k!}, \quad k = 0, 1, \ldots, n,$$

of the polynomial p at the point z. The scheme is recursively defined by $b_m^{(-1)} := a_m$, $m = 0, \ldots, n$, and

$$b_0^{(k)} := b_0^{(k-1)}, \quad b_m^{(k)} := z b_{m-1}^{(k)} + b_m^{(k-1)}, \quad m = 1, \ldots, n-k,$$

for $k = 0, \ldots, n$.

Example 6.23 For the polynomial $p(x) := x^3 - x^2 + 3x - 5$ the Horner scheme

z	1	-1	3	- 5
2	1	1	5	5
2	1	3	11	
2	1	5		
2	1			

for $z = 2$ leads to $p(2) = 5$, $p'(2) = 11$, $p''(2) = 10$, $p'''(2) = 6$. □

We continue by outlining how to compute all the zeros of a polynomial p of degree n with real coefficients. We first assume that p has only simple real zeros and proceed as follows:

1. Either from analytic considerations or by plotting a graph of the polynomial we obtain a rough estimate of the location of the zeros $z_n < z_{n-1} < \cdots < z_2 < z_1$.
2. Starting with some $x_0 > z_1$, by Newton iteration we compute the largest zero z_1. The global convergence of Newton's method in this case follows from monotonicity arguments (see Problem 6.13).
3. By the Horner scheme we divide p by the linear factor $x - z_1$ and carry out step two for the reduced polynomial to compute z_2. Repeating this procedure, we successively obtain approximations for all zeros.
4. In order to improve the accuracy, for all zeros Newton's method is applied to the full polynomial p with the starting points of the iteration given by the approximations obtained in step three.

Now we consider the case of multiple real zeros. If z is a zero of order m, then we can write

$$p(x) = (x - z)^m q(x), \tag{6.12}$$

where the polynomial q of degree $n - m$ has a value $q(z) \neq 0$. To see the effect of (6.12) on Newton's method we consider it as a fixed-point iteration $x_{\nu+1} := g(x_\nu)$ with g defined by

$$g(x) := x - \frac{p(x)}{p'(x)}\,.$$

Using (6.12), by elementary differentiation we obtain

$$g'(z) = 1 - \frac{1}{m}\,.$$

Therefore, by Theorem 6.2, at a multiple zero Newton's method is locally convergent. Obviously, the convergence at a multiple zero is only linear. However, one can modify Newton's method for multiple zeros such that the quadratic convergence is preserved (see Problem 6.14).

For finding complex zeros, in principle one can apply Newton's method in $\mathbb{C}$. For this one has to keep in mind that for polynomials with real coefficients, the starting values need to be complex, since otherwise Newton's method would produce only real approximations. For the conjugate complex zeros of a polynomial with real coefficients *Bairstow's method* avoids working in the complex plane by using the fact that for two conjugate zeros, the product of the linear factors $(x - z)(x - \bar{z})$ is a polynomial of degree two with real coefficients. The basic idea is to write the polynomial p of degree n in the form

$$p(x) = (x^2 - ux - v)q(x) + a(x - u) + b,$$

where q is a polynomial of degree $n-2$, and a and b are constants depending on $u, v \in \mathbb{R}$. The factor $x^2 - ux - v$ corresponds to two conjugate complex zeros of p if the pair u, v solves the nonlinear system $a(u, v) = 0$, $b(u, v) = 0$. The latter can be solved by Newton's method, and once the solution u, v is known, the two zeros of p are obtained by solving the quadratic equation $x^2 - ux - v = 0$.

We conclude this section with some consideration of the question of stability. In particular, we show that the zeros of polynomials can be quite sensitive to small changes in the coefficients even if all the zeros are simple and well separated from each other.

Let p and q be polynomials of degree n and assume that z_0 is a simple zero of p. Consider the perturbed polynomial

$$p(\cdot\,, \varepsilon) := p + \varepsilon q,$$

where ε is small. Using the theory of functions of a complex variable, it can be shown that in a neighborhood of $\varepsilon = 0$ the zero $z(\varepsilon)$ depends analytically

on the parameter ε. The derivative z' can be obtained by differentiating $p[z(\varepsilon), \varepsilon] = 0$ with respect to ε. This yields

$$\{p'[z(\varepsilon)] + \varepsilon q'[z(\varepsilon)]\} z'(\varepsilon) + q[z(\varepsilon)] = 0,$$

and setting $\varepsilon = 0$, it follows that

$$z'(0) = -\frac{q(z_0)}{p'(z_0)} .$$

Hence, for small ε we have that

$$z(\varepsilon) \approx z_0 - \varepsilon \, \frac{q(z_0)}{p'(z_0)} . \tag{6.13}$$

Example 6.24 The polynomial

$$p(x) := (x-1)(x-2)\cdots(x-10) = x^{10} - 55x^9 + \cdots + 10!$$

has the zeros $1, 2, \ldots, 10$, which are well separated from each other. We perturb the coefficient of x^9 by choosing $q(x) := 55x^9$. Since $p'(10) = 9!$, by (6.13), the zero $z_0 = 10$ of the polynomial p is perturbed into

$$10 - \frac{55 \cdot 10^9}{9!} \varepsilon \approx 10 - 1.5 \cdot 10^5 \varepsilon.$$

This illustrates that finding the zeros of p is an ill-conditioned problem and that a reliable approximation of the zeros is impossible. □

6.4 Least Squares Problems

Quite often the problem of solving a system of nonlinear equations may be replaced by an equivalent problem of minimizing a function and vice versa. We illustrate this by introducing the *Levenberg–Marquardt method* as one of the most effective procedures for solving nonlinear least squares problems.

Let $g : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function and consider the problem of minimizing g. Let x_0 be an approximation for a local minimum of g. In a neighborhood of x_0, by Taylor's formula we may approximate

$$g(x) \approx g(x_0) + (x - x_0)^T \operatorname{grad} g(x_0) + \frac{1}{2}\,(x - x_0)^T g''(x_0)(x - x_0), \tag{6.14}$$

where

$$g''(x) = \left(\frac{\partial^2 g}{\partial x_j \partial x_k} \right)_{j,k=1,\ldots,n}$$

denotes the *Hessian matrix* of g. Minimizing the quadratic function on the right-hand side of (6.14) yields

$$x_1 = x_0 - [g''(x_0)]^{-1} \operatorname{grad} g(x_0) \tag{6.15}$$

as a new approximation for the minimum of g. We observe that (6.15) obviously coincides with one Newton step for solving the necessary condition $\operatorname{grad} g(x) = 0$ for a local minimum.

However, if (6.14) is only a very poor approximation to g, then we expect the Newton step (6.15) not to be very effective. In this case it is more appropriate to use a so-called *method of steepest descent;* i.e., choose

$$x_1 = x_0 - \lambda M \operatorname{grad} g(x_0) \tag{6.16}$$

as a new approximation. Here M is a positive definite matrix, and the step size $\lambda > 0$ is chosen such that $g(x_1) < g(x_0)$ is satisfied. This can be achieved, since by Taylor's formula we have that

$$g[x_0 - \lambda M \operatorname{grad} g(x_0)] \approx g(x_0) - \lambda [\operatorname{grad} g(x_0)]^T M \operatorname{grad} g(x_0)$$

and M is assumed to be positive definite.

After introducing the vector $y \in \mathbb{R}^n$ and the $n \times n$ matrix A by

$$y_j(x) := -\frac{\partial g}{\partial x_j}(x), \quad a_{jk}(x) := \frac{\partial^2 g}{\partial x_j \partial x_k}(x), \tag{6.17}$$

we can rewrite the Newton iteration (6.15) as the linear system

$$A(x_0)(x_1 - x_0) = y, \tag{6.18}$$

which we have to solve for the difference $x_1 - x_0$. Similarly, one step of the steepest descent (6.16) can be transformed into

$$x_1 - x_0 = \lambda M y. \tag{6.19}$$

Now recall the least squares problem of Example 2.4. In a slight reformulation, this problem consists in minimizing the function

$$g(x) := \sum_{i=1}^{m} [f_i(x) - u_i]^2$$

over some domain D, where the $f_i : D \to \mathbb{R}$ are given functions and the $u_i \in \mathbb{R}$ are given constants for $i = 1, \ldots, m$. We compute the derivatives

$$\frac{\partial g}{\partial x_j}(x) = 2 \sum_{i=1}^{m} [f_i(x) - u_i] \frac{\partial f_i}{\partial x_j}(x)$$

and

$$\frac{\partial^2 g}{\partial x_j \partial x_k}(x) = 2\sum_{i=1}^{m} \left\{ \frac{\partial f_i}{\partial x_j}(x)\, \frac{\partial f_i}{\partial x_k}(x) + [f_i(x) - u_i]\, \frac{\partial^2 f_i}{\partial x_j \partial x_k}(x) \right\}.$$

In this case the matrix (a_{jk}) contains second derivatives of the functions f_i. However, since these derivatives are multiplied by the factor $[f_i(x)-u_i]$, which will become small by minimizing g, it is justified to neglect this term. Note that if Newton's method converges, it always will converge to a zero, even if we do not use the exact Jacobian for the computation, provided that the approximate Jacobian at the limit is nonsingular. Hence, we simplify and replace (6.17) by

$$a_{jk}(x) := 2\sum_{i=1}^{m} \frac{\partial f_i}{\partial x_j}(x)\, \frac{\partial f_i}{\partial x_k}(x) \tag{6.20}$$

and note that $a_{jj}(x) > 0$.

Now the Levenberg–Marquardt method combines (6.18) and (6.19) by first introducing the $n \times n$ matrix $\tilde{A} = (\tilde{a}_{jk})$ with entries

$$\tilde{a}_{jj} = (1+\gamma)a_{jj}, \quad \tilde{a}_{jk} = a_{jk}, \quad j \neq k,$$

where γ is some positive parameter, and then replacing (6.18) and (6.19) by

$$\tilde{A}(x_0)(x_1 - x_0) = y. \tag{6.21}$$

Obviously, for large γ the matrix $\tilde{A}$ will become diagonally dominant, and (6.21) will get close to the steepest descent, with

$$M = \operatorname{diag}\left(\frac{1}{a_{11}}, \dots, \frac{1}{a_{nn}}\right)$$

and $\lambda = 1/\gamma$. For $\gamma \to 0$, on the other hand, (6.21) will turn into the Newton step (6.18). This ability to gradually vary between Newton's method and the steepest descent method is one of the basic features of the Levenberg–Marquardt method, which we describe as follows:

1. Choose an initial guess x_0, some moderately sized value for γ, and a factor α, say $\gamma = 0.001$ and $\alpha = 10$.
2. Solve the linear system (6.21) to obtain x_1.
3. If $g(x_1) > g(x_0)$, then reject x_1 as a new approximation, replace γ by $\alpha\gamma$, and go back and repeat step two.
4. If $g(x_1) < g(x_0)$, then accept x_0 as a new approximation, replace x_0 by x_1 and γ by γ/α, and go back to step two.
5. Terminate when the difference $|g(x_1) - g(x_0)|$ is smaller than some given tolerance.

For a detailed analysis of this method we refer to [44]. For a study of nonlinear optimization methods and their relation to nonlinear systems we refer to [20].

Problems

6.1 Prove *Brouwer's fixed point theorem* in $\mathbb{R}$; i.e., show that if $D \subset \mathbb{R}$ is a closed and bounded interval and if $f : D \to D$ is continuous, then f has a (not necessarily unique) fixed point.

6.2 Draw figures illustrating monotone or alternating divergence of the successive iterations for a fixed point of a function of one variable.

6.3 Show how to solve the equation $\tan x = x$ by successive approximations.

6.4 Show that

$$\lim_{\nu \to \infty} \underbrace{\sqrt{2 + \sqrt{2 + \cdots + \sqrt{2}}}}_{\nu \text{ square roots}} = 2.$$

6.5 Let $D \subset \mathbb{R}$ be an open interval and let $f : D \to D$ be m times continuously differentiable. Under the assumption that the sequence $x_{\nu+1} := f(x_\nu)$ converges to some x in D with $f'(x) = f''(x) = \cdots = f^{(m-1)}(x) = 0$, show that the convergence is of order m.

6.6 Let the sequence (x_ν) in $\mathbb{R}$ converge to x such that $x_\nu \neq x$ for all $\nu \in \mathbb{N}$ and

$$x_{\nu+1} - x = (q + \xi_\nu)(x_\nu - x), \quad \nu = 0, 1, \ldots,$$

where $|q| < 1$ and $\xi_\nu \to 0$, $\nu \to \infty$. Show that

$$y_\nu := x_\nu - \frac{(x_{\nu+1} - x_\nu)^2}{x_{\nu+2} - 2x_{\nu+1} + x_\nu}$$

is well-defined for sufficiently large ν and that

$$\lim_{\nu \to \infty} \frac{y_\nu - x}{x_\nu - x} = 0;$$

i.e., the sequence (y_ν) converges to x more rapidly than the sequence (x_ν). This method for speeding up the convergence of sequences is known as *Aitken's* δ^2*method.*

6.7 Let $D \subset \mathbb{R}$ be an open interval, let $f : D \to \mathbb{R}$ be twice continuously differentiable, and let x be a fixed point of f with $f'(x) \neq 1$. Show that *Steffensen's method*

$$x_{\nu+1} := x_\nu - \frac{[f(x_\nu) - x_\nu]^2}{f[f(x_\nu)] - 2f(x_\nu) + x_\nu}, \quad \nu = 0, 1, \ldots,$$

is locally and quadratically convergent to the fixed point x (see Problem 6.6).

6.8 Discuss Steffensen's method of Problem 6.7 for the fixed point $x = 0$ of the function $f(x) := 2x + x^3$.

6.9 Show that

$$x_{\nu+1} := \frac{x_\nu(x_\nu^2 + 3a)}{3x_\nu^2 + a}, \quad \nu = 0, 1, \ldots,$$

is a method of order three for computing the square root of a positive number a.

6.10 Prove an analogue of Corollary 6.16 for the secant method (6.9).

6.11 Give conditions for monotone convergence of Newton's method for a function of one variable.

6.12 Show that Newton's method for the function $f(x) := x^n - a$, $x > 0$, where $n > 1$ and $a > 0$, converges globally to $a^{1/n}$.

6.13 Assume that the polynomial p with real coefficients has only real zeros and denote the largest zero by z_1. Show that for any initial point x_0 with $x_0 > z_1$ Newton's method converges to z_1.

6.14 Assume that z is a zero of order m of the polynomial p. Show that

$$x_{\nu+1} := x_\nu - m\,\frac{p(x_\nu)}{p'(x_\nu)}\,, \quad \nu = 0, 1, \dots,$$

converges locally and quadratically to the zero z.

6.15 Show that for a nonsingular $n \times n$ matrix A the sequence

$$A_{\nu+1} := A_\nu[2I - AA_\nu], \quad \nu = 0, 1, \dots,$$

converges quadratically to the inverse A^{-1}, provided that $\|I - AA_0\| < 1$.

6.16 Write a computer program for finding n simple zeros of a polynomial of degree n with real coefficients. Use this code for the computation of the zeros of the Laguerre polynomial $L_4(x) = x^4 - 16\,x^3 + 72\,x^2 - 96\,x + 24$.

6.17 Show that for the function $f : (0, \infty) \to \mathbb{R}$ given by

$$f(x) := \frac{\ln 2}{\pi}\,\sin\left(2\pi\,\frac{\ln x}{\ln 2}\right) + 1$$

the Newton iterations starting with $x_0 = 1$ converge and that the limit, however, is not a zero of f.

6.18 The eigenvalue problem $Ax = \lambda x$ for an $n \times n$ matrix A is equivalent to the equation $f(z) = 0$, where $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n \times \mathbb{R}$ is defined by

$$f : \begin{pmatrix} x \\ \lambda \end{pmatrix} \mapsto \begin{pmatrix} Ax - \lambda x \\ xx^T - 1 \end{pmatrix}.$$

Write down Newton's method for this equation.

6.19 Write a computer program for solving a least squares problem by the Levenberg–Marquardt method.

6.20 The set of all points $\zeta \in \mathbb{C}$ for which the fixed point iteration $z_{\nu+1} := z_\nu^2 + \zeta$ starting with $z_0 = 0$ remains bounded is called the *Mandelbrot set.* Write a computer program for visualizing the Mandelbrot set.

7

Matrix Eigenvalue Problems

Many problems in science and engineering lead to eigenvalue problems for matrices. These occur either directly or by discretization of eigenvalue problems for differential or integral operators. In the latter case the size of the matrices will be rather large. It is the purpose of this chapter to introduce some of the main ideas in matrix eigenvalue computations without attempting to be comprehensive. For a more detailed study we refer to [27, 65].

For the numerical computation of matrix eigenvalues we have to distinguish between two groups of methods:

1. In the so-called *direct methods* the eigenvalues are obtained as zeros of the characteristic polynomial.
2. In contrast, *iterative methods* approximate the eigenvalues through a successive approximation procedure without using the characteristic polynomial.

Since, as illustrated in Example 6.24, the computation of zeros of polynomials of high degree tends in general to be ill-conditioned, in practice iterative methods are used almost exclusively. In this chapter we will discuss the two most important methods of this class, namely the *Jacobi method* and the *QR algorithm*. In the last section we will also briefly describe the *Hessenberg method* as an example of a direct method.

A key factor in all eigenvalue computations is the fact that similarity transformations leave the eigenvalues of a matrix invariant; i.e., for a given matrix A the matrices A and $C^{-1}AC$ have the same eigenvalues for all nonsingular matrices C. This can be seen either from the equivalence of

the equations

$$Ax = \lambda x \quad \text{and} \quad (C^{-1}AC)C^{-1}x = \lambda C^{-1}x$$

or from the multiplication theorem for determinants

$$\det(\lambda I - A) = \det[C^{-1}(\lambda I - A)C] = \det(\lambda I - C^{-1}AC);$$

i.e., similar matrices have the same characteristic polynomial. This invariance allows one to transform a given matrix A by a similarity transformation into a matrix of simpler form with the same eigenvalues as A. In particular, the iterative methods successively construct sequences of similar matrices that converge to a diagonal matrix or an upper (or lower) triangular matrix from which the eigenvalues can be read off as the diagonal elements.

7.1 Examples

We begin by illustrating how the discretization of eigenvalue problems for differential operators leads to eigenvalue problems for large matrices.

Example 7.1 The vibrations of a string are modeled by the so-called *wave equation*

$$\frac{\partial^2 w}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 w}{\partial t^2},$$

where $w = w(x,t)$ denotes the vertical elongation and c is the speed of sound in the string. Assuming that the string is clamped at $x = 0$ and $x = 1$, the boundary conditions $w(0,t) = w(1,t) = 0$ must be satisfied for all times t. Obviously, the time-harmonic wave

$$w(x,t) = v(x)e^{i\omega t}$$

with frequency ω solves the wave equation, provided that the space-dependent part v satisfies

$$-v'' = \lambda v \quad \text{on } [0,1],$$

where $\lambda := \omega^2/c^2$. The boundary conditions $w(0,t) = w(1,t) = 0$ are satisfied if v satisfies the boundary conditions

$$v(0) = v(1) = 0.$$

Hence, introducing the linear space

$$U := \{v \in C[0,1] : v \text{ is twice continuously differentiable}, v(0) = v(1) = 0\}$$

and defining the differential operator $D : U \to C[0,1]$ by $D : v \mapsto -v''$, we are led to the eigenvalue problem $Dv = \lambda v$. Elementary calculations

show that the functions $v_m(x) = \sin m\pi x$ are eigenfunctions of D with the eigenvalues $\lambda_m = m^2\pi^2$ for $m = 1, 2, \ldots$. It can be shown that these are the only eigenvalues and eigenfunctions of D.

For discussing an approximate solution we consider the slightly more general differential equation

$$-v'' + pv = \lambda v \quad \text{on } [0, 1]$$

with boundary conditions $v(0) = v(1) = 0$, where $p \in C[0, 1]$ is a given positive function. We can proceed as in Example 2.1 and choose an equidistant mesh $x_j = jh$, $j = 0, \ldots, n+1$, with step size $h = 1/(n+1)$ and $n \in \mathbb{N}$. At the internal grid points x_j, $j = 1, \ldots, n$, we replace the differential quotient by the difference quotient

$$v''(x_j) \approx \frac{1}{h^2}\{v(x_{j+1}) - 2v(x_j) + v(x_{j-1})\}$$

to obtain the system of equations

$$\frac{1}{h^2}\{-v_{j-1} + 2v_j - v_{j+1}\} + p_j v_j = \lambda v_j, \quad j = 1, \ldots, n,$$

for approximate values v_j to the exact solution $v(x_j)$. Here, we have set $p_j := p(x_j)$ for $j = 0, \ldots, n+1$. This system has to be complemented by the two boundary conditions $v_0 = v_{n+1} = 0$. For an abbreviated notation we introduce the $n \times n$ tridiagonal matrix

$$A = \frac{1}{h^2}\begin{pmatrix} 2+h^2p_1 & -1 & & & & \\ -1 & 2+h^2p_2 & -1 & & & \\ & -1 & 2+h^2p_3 & -1 & & \\ & \cdot & \cdot & \cdot & & \\ & & & -1 & 2+h^2p_{n-1} & -1 \\ & & & & -1 & 2+h^2p_n \end{pmatrix}$$

and the vector $u = (v_1, \ldots, v_n)^T$. Then the above system of equations, including the boundary conditions, reads

$$Au = \lambda u;$$

i.e., the eigenvalue problem for the differential operator D is approximated by the eigenvalue problem for the matrix A. □

The important question as to how well the matrix eigenvalues approximate the eigenvalues of the differential operator and whether we have convergence of the eigenvalues as $h \to 0$ is beyond the scope of this book (see Problem 7.2). The example is meant only as an illustration of the fact that eigenvalue problems for large matrices arise through the discretization of eigenvalue problems for ordinary differential operators and also for partial differential operators. In the same spirit, eigenvalue problems for integral operators can be approximated by matrix eigenvalue problems, as indicated in the following example.

Example 7.2 Consider the eigenvalue problem

$$\int_0^1 K(x,y)\varphi(y)\,dy = \lambda\varphi(x), \quad x \in [0,1],$$

for a linear integral operator with continuous kernel K. For the numerical approximation we proceed as in Example 2.3 and approximate the integral by the rectangular rule with equidistant quadrature points $x_k = k/n$ for $k = 1,\dots,n$. If we require the approximated equation to be satisfied only at the grid points, we arrive at the approximating system of equations

$$\frac{1}{n}\sum_{k=1}^{n} K(x_j,x_k)\varphi_k = \lambda\varphi_j, \quad j = 1,\dots,n,$$

for approximate values φ_j to the exact solution $\varphi(x_j)$. Hence, we approximate the eigenvalues of the integral operator by the eigenvalues of the matrix with entries $K(x_j,x_k)/n$. Of course, instead of the rectangular rule any other quadrature rule can be used. A discussion of the convergence of the matrix eigenvalues to the eigenvalues of the integral operator is again beyond the aim of this introduction. □

7.2 Estimates for the Eigenvalues

At this point we urge the reader to recall the basic facts about eigenvalues of matrices, in particular those that were presented in Section 3.4. In the sequel, by $(\cdot\,,\cdot)$ we denote the Euclidean scalar product in $\mathbb{C}^n$ and by $\|\cdot\|_2$ the corresponding Euclidean norm.

The eigenvalues of Hermitian matrices can be characterized by the following maximum principles. These can be used to get some rough estimates for the eigenvalues. Note that for the eigenvalues of Hermitian matrices the geometric and the algebraic multiplicity coincide (see Problem 7.4).

Theorem 7.3 (Rayleigh) *Let A be a Hermitian $n \times n$ matrix with eigenvalues*

$$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$$

(where multiple eigenvalues occur according to their multiplicity) and corresponding orthonormal eigenvectors $x_1, x_2, \dots, x_n$. Then

$$\lambda_j = \max_{\substack{x\in V_j\\ x\neq 0}} \frac{(Ax,x)}{(x,x)}, \quad j = 1,\dots,n,$$

where the subspaces $V_1,\dots,V_n$ are defined by $V_1 := \mathbb{C}^n$ and

$$V_j := \{x \in \mathbb{C}^n : (x,x_k) = 0,\ k = 1,\dots,j-1\}, \quad j = 2,\dots,n.$$

Proof. Let $x \in V_j$ with $x \neq 0$. Then

$$x = \sum_{k=j}^{n} (x, x_k) x_k \quad \text{and} \quad \sum_{k=j}^{n} |(x, x_k)|^2 = (x, x).$$

Hence

$$Ax = \sum_{k=j}^{n} \lambda_k (x, x_k) x_k$$

and

$$(Ax, x) = \sum_{k=j}^{n} \lambda_k |(x, x_k)|^2 \leq \lambda_j \sum_{k=j}^{n} |(x, x_k)|^2 = \lambda_j (x, x).$$

This implies

$$\sup_{\substack{x \in V_j \\ x \neq 0}} \frac{(Ax, x)}{(x, x)} \leq \lambda_j,$$

and the statement follows from $(Ax_j, x_j) = \lambda_j$ and $x_j \in V_j$. □

This maximum principle can be used in a simple manner to obtain lower bounds for the largest eigenvalue of Hermitian matrices. For the matrix

$$A = \begin{pmatrix} 1 & 3 & 2 \\ 3 & 5 & 1 \\ 2 & 1 & 4 \end{pmatrix},$$

by using $x = (1, 1, 1)^T$ we find the estimate $\lambda_1 \geq 7.33$ as compared to the exact eigenvalue $\lambda_1 = 7.58\ldots$. Using $x = (1, 2, 1)^T$ leads to the estimate $\lambda_1 \geq 7.50$.

Using Rayleigh's principle to obtain bounds for the smaller eigenvalues requires the knowledge of the eigenvectors for the preceding larger eigenvalues. This problem is circumvented in the following minimum maximum principle.

Theorem 7.4 (Courant) *Let A be a Hermitian $n \times n$ matrix with eigenvalues*

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

(where multiple eigenvalues occur according to their multiplicity). Then

$$\lambda_j = \min_{U_j \in M_j} \max_{\substack{x \in U_j \\ x \neq 0}} \frac{(Ax, x)}{(x, x)}, \quad j = 1, \ldots, n,$$

where M_j denotes the set of all subspaces $U_j \subset \mathbb{C}^n$ of dimension $n + 1 - j$.

Proof. First we note that because of

$$\sup_{\substack{x \in U_j \\ x \neq 0}} \frac{(Ax, x)}{(x, x)} = \sup_{\substack{x \in U_j \\ (x,x)=1}} (Ax, x)$$

and the continuity of the function $x \mapsto (Ax, x)$, the supremum is attained; i.e., the maximum exists.

By $x_1, x_2, \dots, x_n$ we denote orthonormal eigenvectors corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. First, we show that for a given subspace U_j of dimension $n + 1 - j$ there exists a vector $x \in U_j$ such that

$$(x, x_k) = 0, \quad k = j + 1, \dots, n. \tag{7.1}$$

Let $z_1, \dots, z_{n+1-j}$ be a basis of U_j. Then we can represent each $x \in U_j$ by

$$x = \sum_{i=1}^{n+1-j} a_i z_i. \tag{7.2}$$

In order to guarantee (7.1), the $n + 1 - j$ coefficients $a_1, \dots, a_{n+1-j}$ must satisfy the $n - j$ linear equations

$$\sum_{i=1}^{n+1-j} a_i (z_i, x_k) = 0, \quad k = j + 1, \dots, n.$$

This underdetermined system always has a nontrivial solution. For the corresponding x given by (7.2) we have $x \neq 0$, and from

$$x = \sum_{k=1}^{j} (x, x_k) x_k$$

we obtain that

$$(Ax, x) = \sum_{k=1}^{j} \lambda_k |(x, x_k)|^2 \geq \lambda_j \sum_{k=1}^{j} |(x, x_k)|^2 = \lambda_j (x, x),$$

whence

$$\max_{\substack{x \in U_j \\ x \neq 0}} \frac{(Ax, x)}{(x, x)} \geq \lambda_j$$

follows.

On the other hand, for the subspace

$$U_j = \{x \in \mathbb{C}^n : (x, x_k) = 0, \ k = 1, \dots, j - 1\}$$

of dimension $n + 1 - j$, by Theorem 7.3 we have the equality

$$\max_{\substack{x \in U_j \\ x \neq 0}} \frac{(Ax, x)}{(x, x)} = \lambda_j,$$

and the proof is finished. □

Corollary 7.5 *Let A and B be two Hermitian $n \times n$ matrices with eigenvalues $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_n(A)$ and $\lambda_1(B) \geq \lambda_2(B) \geq \cdots \geq \lambda_n(B)$. Then*

$$|\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|, \quad j = 1, \ldots, n,$$

for any norm $\|\cdot\|$ on $\mathbb{C}^n$.

Proof. From the Cauchy–Schwarz inequality we have that

$$(Ax - Bx, x) \leq \|(A - B)x\|_2 \, \|x\|_2 \leq \|A - B\|_2 \, \|x\|_2^2$$

and hence

$$(Ax, x) \leq (Bx, x) + \|A - B\|_2 \, \|x\|_2^2.$$

By the Courant minimum maximum principle of Theorem 7.4 this implies

$$\lambda_j(A) \leq \lambda_j(B) + \|A - B\|_2, \quad j = 1 \ldots, n.$$

Interchanging the roles of A and B, we also have that

$$\lambda_j(B) \leq \lambda_j(A) + \|B - A\|_2, \quad j = 1 \ldots, n,$$

and therefore

$$|\lambda_j(A) - \lambda_j(B)| \leq \|A - B\|_2, \quad j = 1, \ldots, n.$$

Now the statement follows from

$$\|A - B\|_2 = \rho(A - B) \leq \|A - B\|,$$

which is a consequence of Theorems 3.31 and 3.32. □

Corollary 7.6 *For the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ of a Hermitian $n \times n$ matrix $A = (a_{jk})$ we have that*

$$|\lambda_i - a'_{ii}|^2 \leq \sum_{\substack{j,k=1 \\ j \neq k}}^{n} |a_{jk}|^2, \quad i = 1, \ldots, n,$$

where the elements $a'_{11}, \ldots, a'_{nn}$ represent a permutation of the diagonal elements $a_{11}, \ldots, a_{nn}$ of A such that $a'_{11} \geq a'_{22} \geq \cdots \geq a'_{nn}$.

Proof. Use $B = \operatorname{diag}(a'_{jj})$ and $\|\cdot\| = \|\cdot\|_2$ in the preceding corollary. □

We conclude this section with an extension of the above results to general matrices that gives a rough estimate as to where in $\mathbb{C}$ the eigenvalues are located.

Theorem 7.7 (Gerschgorin) *Let $A = (a_{jk})$ be a complex $n \times n$ matrix and define the disks*

$$G_j := \left\{ \lambda \in \mathbb{C} : |\lambda - a_{jj}| \le \sum_{\substack{k=1 \\ k \ne j}}^{n} |a_{jk}| \right\}, \quad j = 1, \dots, n,$$

and

$$G_j^* := \left\{ \lambda \in \mathbb{C} : |\lambda - a_{jj}| \le \sum_{\substack{k=1 \\ k \ne j}}^{n} |a_{kj}| \right\}, \quad j = 1, \dots, n.$$

Then the eigenvalues λ of A satisfy

$$\lambda \in \bigcup_{j=1}^{n} G_j \cap \bigcup_{j=1}^{n} G_j^*.$$

Proof. Assume that $Ax = \lambda x$ and $\|x\|_\infty = 1$, and for $x = (x_1, \dots, x_n)^T$ choose j such that $|x_j| = \|x\|_\infty = 1$. Then

$$|\lambda - a_{jj}| = |(\lambda - a_{jj})x_j| = \left| \sum_{\substack{k=1 \\ k \ne j}}^{n} a_{jk} x_k \right| \le \sum_{\substack{k=1 \\ k \ne j}}^{n} |a_{jk}|,$$

and therefore

$$\lambda \in \bigcup_{j=1}^{n} G_j.$$

Since the eigenvalues of A^* are the complex conjugate of the eigenvalues of A (see Problem 7.3) we also have that

$$\lambda \in \bigcup_{j=1}^{n} G_j^*,$$

and the theorem is proven. □

7.3 The Jacobi Method

The method described in this section was discovered by Jacobi in 1846 and can be used to iteratively compute all the eigenvalues and eigenvectors of real symmetric matrices.

Lemma 7.8 *The* Frobenius norm

$$\|A\|_F := \left(\sum_{j,k=1}^{n} |a_{jk}|^2 \right)^{1/2}$$

of an $n \times n$ matrix $A = (a_{jk})$ is invariant with respect to unitary transformations.

Proof. The *trace*

$$\operatorname{tr} A := \sum_{j=1}^{n} a_{jj}$$

of a matrix A is commutative; i.e., $\operatorname{tr} AB = \operatorname{tr} BA$. This follows from

$$\sum_{j=1}^{n} (AB)_{jj} = \sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk} b_{kj} = \sum_{k=1}^{n} \sum_{j=1}^{n} b_{kj} a_{jk} = \sum_{k=1}^{n} (BA)_{kk}.$$

In particular, we have that

$$\operatorname{tr} AA^* = \sum_{j=1}^{n} \sum_{k=1}^{n} a_{jk} a_{kj}^* = \sum_{j=1}^{n} \sum_{k=1}^{n} |a_{jk}|^2.$$

Therefore, for each unitary matrix Q it follows that

$$\|Q^*AQ\|_F^2 = \operatorname{tr}(Q^*AQQ^*A^*Q) = \operatorname{tr}(Q^*AA^*Q) = \operatorname{tr}(AA^*QQ^*) = \|A\|_F^2,$$

and the lemma is proven. □

Corollary 7.9 *The eigenvalues of an $n \times n$ matrix A (counted repeatedly according to their algebraic multiplicity) satisfy* Schur's inequality

$$\sum_{j=1}^{n} |\lambda_j|^2 \le \|A\|_F^2.$$

Equality holds if and only if the matrix A is normal, *i.e., if $AA^* = A^*A$.*

Proof. By Theorem 3.27 there exists a unitary matrix Q such that $R := Q^*AQ$ is an upper triangular matrix. Hence

$$\|A\|_F^2 = \|R\|_F^2 = \sum_{j=1}^{n} |\lambda_j|^2 + \sum_{j=1}^{n} \sum_{k=j+1}^{n} |r_{jk}|^2, \tag{7.3}$$

since the diagonal elements of $R = (r_{jk})$ coincide with the eigenvalues of the similar matrices R and A. Now Schur's inequality follows immediately from (7.3).

For the discussion of the case of equality, we first note that any unitary transformation of a normal matrix is again normal. This is a consequence of the identity

$$Q^*AQ(Q^*AQ)^* - (Q^*AQ)^*Q^*AQ = Q^*(AA^* - A^*A)Q.$$

If equality holds in Schur's inequality, then (7.3) implies that R is a diagonal matrix. Hence R, and therefore A, is normal.

Conversely, if A is normal, then the upper triangular matrix R must also be normal. Now, from

$$(RR^*)_{jj} = \sum_{k=1}^{n} r_{jk} r_{kj}^* = \sum_{k=j}^{n} |r_{jk}|^2$$

and

$$(R^*R)_{jj} = \sum_{k=1}^{n} r_{jk}^* r_{kj} = \sum_{k=1}^{j} |r_{kj}|^2$$

we conclude that

$$\sum_{k=j}^{n} |r_{jk}|^2 = \sum_{k=1}^{j} |r_{kj}|^2, \quad j = 1, \ldots, n.$$

This implies $r_{jk} = 0$ for $j < k$, i.e., R is a diagonal matrix, and from (7.3) we deduce that equality holds in Schur's inequality if A is normal. □

For any $n \times n$ matrix $A = (a_{jk})$ we introduce the quantity

$$N(A) := \left(\sum_{\substack{j,k=1 \\ j \neq k}}^{n} |a_{jk}|^2 \right)^{1/2} \tag{7.4}$$

as a measure for the deviation of A from a diagonal matrix.

Lemma 7.10 *Normal matrices A satisfy*

$$\sum_{j=1}^{n} |\lambda_j|^2 = \sum_{j=1}^{n} |a_{jj}|^2 + [N(A)]^2.$$

Proof. This follows from Corollary 7.9. □

The main idea of the Jacobi method for real symmetric matrices is to successively reduce $N(A)$ by elementary plane rotation matrices such that in the limit the matrix becomes diagonal (with the eigenvalues as diagonal entries).

Lemma 7.11 *For each pair $j < k$ and each $\varphi \in \mathbb{R}$ the matrix*

$$U = \begin{pmatrix} 1 & & & & & & \\ & \cdot & & & & & \\ & & \cos\varphi & & -\sin\varphi & & \\ & & & \cdot & & & \\ & & \sin\varphi & & \cos\varphi & & \\ & & & & & \cdot & \\ & & & & & & 1 \end{pmatrix},$$

which coincides with the identity matrix except for $u_{jj} = u_{kk} = \cos\varphi$ and $u_{kj} = -u_{jk} = \sin\varphi$ (and which describes a rotation in the $x_j x_k$-plane) is unitary.

Proof. This follows from

$$\begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and

$$\begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix} \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Lemma 7.12 *Let A be a real symmetric matrix and let U be the unitary matrix of Lemma 7.11. Then $B = U^*AU$ is also real and symmetric and has the entries*

$$\begin{aligned}
b_{jj} &= a_{jj}\cos^2\varphi + a_{jk}\sin 2\varphi + a_{kk}\sin^2\varphi, \\
b_{kk} &= a_{jj}\sin^2\varphi - a_{jk}\sin 2\varphi + a_{kk}\cos^2\varphi, \\
b_{jk} &= b_{kj} = a_{jk}\cos 2\varphi + \frac{1}{2}\,(a_{kk} - a_{jj})\sin 2\varphi, \\
b_{ij} &= b_{ji} = a_{ij}\cos\varphi + a_{ik}\sin\varphi, \quad i \neq j, k, \\
b_{ik} &= b_{ki} = -a_{ij}\sin\varphi + a_{ik}\cos\varphi, \quad i \neq j, k, \\
b_{il} &= a_{il}, \quad i, l \neq j, k;
\end{aligned}$$

i.e., the matrix B differs from A only in the jth and kth rows and columns.

Proof. The matrix B is real, since A and U are real, and it is symmetric, since the unitary transformation of a Hermitian matrix is again Hermitian. Elementary calculations show that

$$\begin{pmatrix} \cos\varphi & \sin\varphi \\ -\sin\varphi & \cos\varphi \end{pmatrix} \begin{pmatrix} a_{jj} & a_{jk} \\ a_{kj} & a_{kk} \end{pmatrix} \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} = \begin{pmatrix} b_{jj} & b_{jk} \\ b_{kj} & b_{kk} \end{pmatrix}$$

with b_{jj}, b_{jk}, b_{kj}, and b_{kk} as stated in the theorem. For $i \neq j, k$ we have that

$$b_{ij} = \sum_{r,s=1}^{n} u_{is}^* a_{sr} u_{rj} = a_{ij} u_{jj} + a_{ik} u_{kj} = a_{ij} \cos\varphi + a_{ik} \sin\varphi$$

and

$$b_{ik} = \sum_{r,s=1}^{n} u_{is}^* a_{sr} u_{rk} = a_{ij} u_{jk} + a_{ik} u_{kk} = -a_{ij} \sin\varphi + a_{ik} \cos\varphi.$$

Finally, we have

$$b_{il} = \sum_{r,s=1}^{n} u_{is}^* a_{sr} u_{rl} = a_{il}$$

for $i, l \neq j, k$. □

Lemma 7.13 *For*

$$\tan 2\varphi = \frac{2a_{jk}}{a_{jj} - a_{kk}}, \qquad a_{jj} \neq a_{kk},$$

$$\varphi = \frac{\pi}{4}, \qquad a_{jj} = a_{kk},$$

the transformation of Lemma 7.12 annihilates the elements

$$b_{jk} = b_{kj} = 0$$

and reduces the off-diagonal elements according to

$$[N(B)]^2 = [N(A)]^2 - 2a_{jk}^2.$$

Proof. $b_{jk} = b_{kj} = 0$ follows immediately from Lemma 7.12. Applying Lemma 7.8 to the matrices

$$\begin{pmatrix} a_{jj} & a_{jk} \\ a_{kj} & a_{kk} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} b_{jj} & b_{jk} \\ b_{kj} & b_{kk} \end{pmatrix}$$

yields

$$a_{jj}^2 + 2a_{jk}^2 + a_{kk}^2 = b_{jj}^2 + b_{kk}^2.$$

From this, with the aid of Lemmas 7.8 and 7.12 we find that

$$[N(B)]^2 = \|B\|_F^2 - \sum_{i=1}^{n} b_{ii}^2 = \|A\|_F^2 - \sum_{i=1}^{n} b_{ii}^2$$

$$= [N(A)]^2 + \sum_{i=1}^{n} (a_{ii}^2 - b_{ii}^2) = [N(A)]^2 - 2a_{jk}^2,$$

which completes the proof. □

Note that the quantities required for the computation of the elements of the transformed matrix can be obtained by the trigonometric identities

$$\cos 2\varphi = \frac{1}{\sqrt{1+\tan^2 2\varphi}}\,,$$

$$\cos\varphi = \sqrt{\frac{1}{2}\,(1+\cos 2\varphi)}, \quad \sin\varphi = \sqrt{\frac{1}{2}\,(1-\cos 2\varphi)}.$$

The sign of the root in the expression for $\sin\varphi$ has to be chosen such that it coincides with the sign of $\tan 2\varphi$.

The *classical Jacobi method* generates a sequence (A_ν) of similar matrices by starting with the given matrix $A_0 := A$ and choosing the unitary transformation at the νth step according to Lemma 7.13 such that the nondiagonal element of $A_{\nu-1}$ with largest absolute value is annihilated. It is obvious that the elements annihilated in one step of the Jacobi iteration, in general, do not remain zero during subsequent steps. However, we can establish the following convergence result.

Theorem 7.14 *The classical Jacobi method converges; i.e., the sequence* (A_ν) *converges to a diagonal matrix with the eigenvalues of* A *as diagonal elements.*

Proof. For one step of the Jacobi method, from

$$[N(A)]^2 \le (n^2-n) \max_{\substack{i,l=1,\dots,n\\ i\neq l}} a_{il}^2$$

we obtain that

$$a_{jk}^2 \ge \frac{[N(A)]^2}{n(n-1)}$$

for the nondiagonal element a_{jk} with largest modulus. Hence, from Lemma 7.13 we deduce that

$$[N(B)]^2 = [N(A)]^2 - 2a_{jk}^2 \le q^2[N(A)]^2,$$

where

$$q := \left(1 - \frac{2}{n(n-1)}\right)^{1/2}.$$

For the sequence (A_ν) this implies that

$$N(A_\nu) \le q^\nu N(A_0)$$

for all $\nu \in \mathbb{N}$, whence $N(A_\nu) \to 0$, $\nu \to \infty$, since $q < 1$. □

Note that for large n the value of q is close to one, indicating a slow convergence of the Jacobi method. Writing $A_\nu = (a_{jk,\nu})$ by Corollary 7.6 we have the a posteriori error estimate

$$|\lambda_j - a_{jj,\nu}| \le N(A_\nu), \quad j = 1, \dots, n,$$

after performing ν steps of the Jacobi method. Further error estimates can be derived from Gerschgorin's Theorem 7.7.

Approximations to the eigenvectors can be obtained by successively multiplying the unitary transformations of each step. We have $A_\nu = Q_\nu^* A Q_\nu$, where $Q_\nu = U_1 \cdots U_\nu$ is the product of the elementary unitary transformations for each step. From

$$A_\nu \approx D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$$

it follows that $AQ_\nu \approx Q_\nu D$. Hence the columns $Q_\nu = (u_1, \dots, u_n)$ of Q_ν satisfy $Au_j \approx \lambda_j u_j$ for $j = 1, \dots, n$; i.e., they provide approximations to the eigenvectors.

In each step, the classical Jacobi method requires the determination of the nondiagonal element with largest modulus. In order to reduce the computational costs, in the *cyclic Jacobi method* the nondiagonal elements are annihilated in the order

$$(1,2), \dots, (1,n), (2,3), \dots, (2,n), (3,4), \dots, (n-1,n)$$

independent of their size. Convergence results can also be established for this variant (see [27]).

A further refinement is to choose a constant threshold and to annihilate in each cyclic sweep only those off-diagonal elements that are larger in absolute value than the threshold. Of course, the threshold needs to be lowered after each sweep, i.e., after performing a full cycle. For details we refer to [48, 65].

Example 7.15 For the matrix

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$$

the first six transformed matrices for the classical Jacobi method are given by

$$A_1 = \begin{pmatrix} 1.0000 & 0.0000 & -0.7071 \\ 0.0000 & 3.0000 & -0.7071 \\ -0.7071 & -0.7071 & 2.0000 \end{pmatrix},$$

$$A_2 = \begin{pmatrix} 0.6340 & -0.3251 & 0.0000 \\ -0.3251 & 3.0000 & -0.6280 \\ 0.0000 & -0.6280 & 2.3660 \end{pmatrix},$$

$$A_3 = \begin{pmatrix} 0.6340 & -0.2768 & -0.1704 \\ -0.2768 & 3.3864 & 0.0000 \\ -0.1704 & 0.0000 & 1.9796 \end{pmatrix},$$

$$A_4 = \begin{pmatrix} 0.6064 & 0.0000 & -0.1695 \\ 0.0000 & 3.4140 & 0.0169 \\ -0.1695 & 0.0169 & 1.9796 \end{pmatrix},$$

$$A_5 = \begin{pmatrix} 0.5858 & 0.0020 & 0.0000 \\ 0.0020 & 3.4140 & 0.0168 \\ 0.0000 & 0.0168 & 2.0002 \end{pmatrix},$$

$$A_6 = \begin{pmatrix} 0.5858 & 0.0020 & -0.0000 \\ 0.0020 & 3.4142 & 0.0000 \\ -0.0000 & 0.0000 & 2.0000 \end{pmatrix}.$$

The exact eigenvalues of A are $\lambda_1 = 2 + \sqrt{2}$, $\lambda_2 = 2$, $\lambda_3 = 2 - \sqrt{2}$. □

7.4 The QR Algorithm

The *QR algorithm* was suggested by Francis in 1961 and is an iterative method for computing all eigenvalues and eigenvectors for arbitrary complex matrices. In applications, it is the most commonly used method for eigenvalue computations. Our presentation of the QR algorithm follows [62].

For motivation we first consider the *power method* introduced by von Mises in 1929 for finding the eigenvalue with largest modulus.

Definition 7.16 *A matrix A is called* diagonalizable *if there exists a nonsingular matrix C such that $C^{-1}AC$ is a diagonal matrix; i.e., A is similar to a diagonal matrix.*

Theorem 7.17 *An $n \times n$ matrix A is diagonalizable if and only if it has n linearly independent eigenvectors.*

Proof. Assume that $C^{-1}AC = D$, where $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, is diagonal. Then $De_j = \lambda_j e_j$, $j = 1, \ldots, n$, with the canonical orthonormal basis $e_1, \ldots, e_n$ of $\mathbb{C}^n$. This implies that the vectors $x_j := Ce_j$, $j = 1, \ldots, n$, are eigenvectors of A, since

$$Ax_j = ACe_j = CDe_j = C\lambda_j e_j = \lambda_j x_j.$$

The vectors $x_1, \dots, x_n$ are linearly independent because C is nonsingular and the $e_1, \dots, e_n$ are linearly independent.

Conversely, assume that $x_1, \dots, x_n$ are n linearly independent eigenvectors of A for the eigenvalues $\lambda_1, \dots, \lambda_n$. Then the matrix $C = (x_1, \dots, x_n)$ formed by the eigenvectors as columns is nonsingular, and we have that

$$AC = (Ax_1, \dots, Ax_n) = (\lambda_1 x_1, \dots, \lambda_n x_n) = CD,$$

where $D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. Hence $C^{-1}AC = D$. □

We order the eigenvalues of a diagonalizable $n \times n$ matrix A according to their absolute values and assume that

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|;$$

i.e., there is only one eigenvalue of maximal modulus. Starting from an arbitrary vector $v_0 \in \mathbb{C}^n$ we construct the sequence

$$v_\nu = A^\nu v_0, \quad \nu = 1, 2, \dots,$$

by the successive iterations $v_\nu := Av_{\nu-1}$. Note that in order to avoid numerical overflow or underflow we need to scale after each step. Since the n linearly independent eigenvectors $x_1, \dots, x_n$ of A form a basis of $\mathbb{C}^n$, we can represent

$$v_0 = \sum_{k=1}^{n} \alpha_k x_k,$$

whence

$$A^\nu v_0 = \sum_{k=1}^{n} \alpha_k \lambda_k^\nu x_k$$

follows. Scaling after each step by the factor $1/\lambda_1$ leads to

$$\frac{A^\nu v_0}{\lambda_1^\nu} = \sum_{k=1}^{n} \alpha_k \left(\frac{\lambda_k}{\lambda_1} \right)^\nu x_k,$$

and consequently

$$\frac{A^\nu v_0}{\lambda_1^\nu} \to \alpha_1 x_1 \quad \text{and} \quad \frac{\|v_{\nu+1}\|_2}{\|v_\nu\|_2} \to |\lambda_1|$$

as $\nu \to \infty$, provided that $\alpha_1 \neq 0$. Of course, in principle, λ_1 cannot be used as a scaling factor, since it is not known. However, this is irrelevant, since the eigenvector is determined only up to multiplication by a complex constant; i.e., only the direction of the eigenvector is relevant. In practical computations, the condition $\alpha_1 \neq 0$, i.e., $v_0 \notin \operatorname{span}\{x_2, \dots, x_n\}$, will be automatically satisfied through roundoff errors.

The fact that we need to find only the direction of the eigenvectors motivates us to interpret the power method as a successive iteration of subspaces. For

$$S := \mathrm{span}\{v_0\} \quad \text{and} \quad A^\nu S = \mathrm{span}\{A^\nu v_0\}$$

from the above we have that $A^\nu S \to \mathrm{span}\{x_1\}$, $\nu \to \infty$. More generally, we can choose any subspace S of dimension $1 \leq \dim S < n$ and iterate $A^\nu S = \{A^\nu v : v \in S\}$.

Lemma 7.18 *Let A be a diagonalizable $n \times n$ matrix with eigenvalues*

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$$

and corresponding eigenvectors $x_1, x_2, \ldots, x_n$. Assume that for some m with $1 \leq m < n$ we have that $|\lambda_m| > |\lambda_{m+1}|$ and define

$$T := \mathrm{span}\{x_1, \ldots, x_m\} \quad \textit{and} \quad U := \mathrm{span}\{x_{m+1}, \ldots, x_n\}.$$

Further, assume that S is a subspace of $\mathbb{C}^n$ with dimension m satisfying

$$S \cap U = \{0\}.$$

Then the orthogonal projections $P_{A^\nu S}$ and P_T of $\mathbb{C}^n$ onto $A^\nu S$ and T, respectively, satisfy

$$\|P_{A^\nu S} - P_T\|_2 \leq M \left|\frac{\lambda_{m+1}}{\lambda_m}\right|^\nu, \quad \nu \in \mathbb{N},$$

for some constant M; i.e., the subspaces $A^\nu S$ converge to T.

Proof. **1.** First, we show that we can choose a convenient basis for S. Let $y_1, \ldots, y_m$ denote a given basis of S. Then, for $i = 1, \ldots, m$, we can represent

$$y_j = \sum_{k=1}^m b_{jk} x_k + v_j, \tag{7.5}$$

where $v_j \in U$. We prove that the $m \times m$ matrix $B = (b_{jk})$ is nonsingular. To accomplish this, assume that $\alpha_1, \ldots, \alpha_m$ solve the homogeneous adjoint system

$$\sum_{j=1}^m b_{jk} \alpha_j = 0, \quad k = 1, \ldots, m.$$

Then from (7.5) it follows that

$$\sum_{j=1}^m \alpha_j y_j = \sum_{j=1}^m \alpha_j v_j,$$

and from this, with the aid of $S \cap U = \{0\}$ and the linear independence of the y_j, we conclude that $\alpha_1 = \cdots = \alpha_m = 0$. Hence, B indeed is nonsingular.

We denote the entries of the inverse of B by $B^{-1} = (c_{jk})$. Then

$$z_j := \sum_{k=1}^{m} c_{jk} y_k, \quad j = 1, \ldots, m,$$

defines a new basis for S of the form

$$z_j = x_j + u_j,$$

where $u_j \in U$ for $j = 1, \ldots, m$. Because of

$$A^\nu z_j = \lambda_j^\nu x_j + A^\nu u_j, \quad j = 1, \ldots, m,$$

the linearly independent vectors

$$w_{j\nu} := \frac{A^\nu z_j}{\lambda_j^\nu} = x_j + \frac{A^\nu u_j}{\lambda_j^\nu}, \quad j = 1, \ldots, m,$$

form a basis of $A^\nu S$. Since we can represent any $u \in U$ in the form

$$u = \sum_{k=m+1}^{n} \alpha_k x_k,$$

from

$$A^\nu u = \sum_{k=m+1}^{n} \alpha_k \lambda_k^\nu x_k$$

we conclude that there exists a constant $L > 0$ such that

$$\|w_{j\nu} - x_j\|_2 \le L \left| \frac{\lambda_{m+1}}{\lambda_m} \right|^\nu, \quad j = 1, \ldots, m, \tag{7.6}$$

for all $\nu \in \mathbb{N}$.

2. By Corollary 3.53, the orthogonal projection of an element $\eta \in \mathbb{C}^n$ onto the subspace T is given by

$$P_T \eta = \sum_{k=1}^{m} \alpha_k x_k, \tag{7.7}$$

where the coefficients $\alpha_1, \ldots, \alpha_m$ solve the normal equations

$$\sum_{k=1}^{m} \alpha_k (x_k, x_j) = (\eta, x_j), \quad j = 1, \ldots, m. \tag{7.8}$$

Analogously, we have

$$P_{A^\nu S} \eta = \sum_{k=1}^{m} \beta_{k\nu} w_{k\nu} \tag{7.9}$$

and

$$\sum_{k=1}^{m} \beta_{k\nu}(w_{k\nu}, w_{j\nu}) = (\eta, w_{j\nu}), \quad j = 1, \ldots, m. \tag{7.10}$$

We denote the $m \times m$ matrices of the linear systems (7.8) and (7.10) by X and W_ν, respectively. Then, with the aid of the Cauchy–Schwarz inequality, (7.6) implies that

$$\|W_\nu - X\|_2 \leq C_1 \left| \frac{\lambda_{m+1}}{\lambda_m} \right|^\nu, \quad \nu \in \mathbb{N}, \tag{7.11}$$

for some constant C_1. We denote the right-hand sides of (7.8) and (7.10) by a and b_ν, respectively. Again from (7.6) and the Cauchy–Schwarz inequality we have that

$$\|b_\nu - a\|_2 \leq C_2 \left| \frac{\lambda_{m+1}}{\lambda_m} \right|^\nu \|\eta\|_2, \quad \nu \in \mathbb{N}, \tag{7.12}$$

for some constant C_2. Now, considering the linear system (7.10) as a perturbation of (7.8), from Theorem 5.3 we can conclude that

$$\|\beta_\nu - \alpha\|_2 \leq C_3 \left| \frac{\lambda_{m+1}}{\lambda_m} \right|^\nu \|\alpha\|_2, \quad \nu \in \mathbb{N}, \tag{7.13}$$

for the vectors $\alpha = (\alpha_1, \ldots, \alpha_m)^T$ and $\beta_\nu = (\beta_{1\nu}, \ldots, \beta_{m\nu})^T$ and some constant C_3. From (7.7) and (7.9), using (7.6), (7.13), and the triangle inequality, the assertion of the lemma follows. □

The subspace T of Lemma 7.18 is invariant with respect to A; i.e., $A(T) = T$. By a knowledge of invariant subspaces the eigenvalue problem for the full matrix A can be reduced to eigenvalue problems for two smaller matrices. Assume that

$$P = (P_1, P_2)$$

is a unitary matrix such that its first m columns represented by the matrix P_1 form a basis of T. Then $P_2^* A P_1 = 0$, since T is invariant with respect to A, and $P_2^* P_1 = 0$. Therefore, the unitary transformation yields

$$P^*AP = \begin{pmatrix} P_1^*AP_1 & P_1^*AP_2 \\ P_2^*AP_1 & P_2^*AP_2 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix};$$

i.e., the eigenvalue problem for A is reduced to two smaller eigenvalue problems for the $m \times m$ matrix A_{11} and the $(n-m) \times (n-m)$ matrix A_{22}.

The successive iterations of Lemma 7.18 yield only approximations $A^\nu S$ to the invariant subspace T. However, if

$$Q_\nu = (Q_{1\nu}, Q_{2\nu})$$

denotes a unitary matrix such that its first m columns represented by the matrix $Q_{1\nu}$ form a basis of $A^\nu S$, then for

$$Q_\nu^* A Q_\nu = \begin{pmatrix} B_{11,\nu} & B_{12,\nu} \\ B_{21,\nu} & B_{22,\nu} \end{pmatrix}$$

we expect that $B_{21,\nu} \to 0$, $\nu \to \infty$. Before we can establish this result we need to investigate further the iteration of subspaces.

Choose a basis $y_1, \dots, y_n$ of $\mathbb{C}^n$ and consider the subspaces

$$S_m := \operatorname{span}\{y_1, \dots, y_m\}, \quad m = 1, \dots, n-1.$$

For a simultaneous iteration of all the subspaces $A^\nu S_m$ it clearly suffices to iterate the basis vectors $A^\nu y_1, \dots, A^\nu y_n$. If the assumptions of Lemma 7.18 are satisfied for each $m = 1, \dots, n-1$, then

$$A^\nu S_m \to T_m := \operatorname{span}\{x_1, \dots, x_m\}, \quad \nu \to \infty,$$

for $m = 1, \dots, n-1$. Hence we expect to be able to construct unitary matrices Q_ν such that $Q_\nu^* A Q_\nu \to R$, $\nu \to \infty$, where R is an upper triangular matrix that is similar to A.

For the actual computation two difficulties arise. Firstly, the iterated vectors have to be scaled in order to avoid numerical overflow or underflow. Secondly, by Theorem 7.17, as $\nu \to \infty$ each of the n sequences $(A^\nu y_1), \dots, (A^\nu y_n)$ will converge to the subspace $\operatorname{span}\{x_1\}$ spanned by the eigenvector for the eigenvalue λ_1 with largest modulus. Hence, for large ν the vectors $A^\nu y_1, \dots, A^\nu y_n$ will be almost collinear; i.e., the basis elements $A^\nu y_1, \dots, A^\nu y_n$ are almost linearly dependent and therefore ill-conditioned for spanning the iterated subspaces.

Both these difficulties can be remedied by orthonormalizing the basis after each step. Assume that $q_{1\nu}, \dots, q_{n\nu}$ are orthonormal vectors such that

$$A^\nu S_m = \operatorname{span}\{q_{1\nu}, \dots, q_{m\nu}\}, \quad m = 1, \dots, n-1.$$

Then we compute $Aq_{1\nu}, \dots, Aq_{n\nu}$ and orthonormalize these vectors from left to right to obtain the vectors $r_{1\nu}, \dots, r_{n\nu}$. This procedure preserves the property

$$\begin{aligned} \operatorname{span}\{r_{1\nu}, \dots, r_{m\nu}\} &= \operatorname{span}\{Aq_{1\nu}, \dots, Aq_{m\nu}\} \\ &= A(\operatorname{span}\{q_{1\nu}, \dots, q_{k\nu}\}) = A^{\nu+1} S_m \end{aligned}$$

for $m = 1, \dots, n-1$.

Theorem 7.19 *Assume that A is a diagonalizable $n \times n$ matrix with eigenvalues*

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$$

and corresponding eigenvectors $x_1, x_2, \ldots, x_n$, *and set*

$$T_m := \operatorname{span}\{x_1, \ldots, x_m\} \quad \textit{and} \quad U_m := \operatorname{span}\{x_{m+1}, \ldots, x_n\}$$

for $m = 1, \ldots, n-1$. *Let* $q_{10}, \ldots, q_{n0}$ *be an orthonormal basis of* $\mathbb{C}^n$ *and let the subspaces*

$$S_m := \operatorname{span}\{q_{10}, \ldots, q_{m0}\}$$

satisfy

$$S_m \cap U_m = \{0\}, \quad m = 1, \ldots, n-1.$$

Assume that for each $\nu \in \mathbb{N}$ *we have constructed an orthonormal system* $q_{1\nu}, \ldots, q_{n\nu}$ *with the property*

$$A^\nu S_m = \operatorname{span}\{q_{1\nu}, \ldots, q_{m\nu}\}, \quad m = 1, \ldots, n-1, \tag{7.14}$$

and define $\tilde{Q}_\nu = (q_{1\nu}, \ldots, q_{n\nu})$. *Then for the sequence of matrices* $A_\nu = (a_{jk,\nu})$ *given by*

$$A_{\nu+1} := \tilde{Q}_\nu^* A \tilde{Q}_\nu \tag{7.15}$$

we have convergence:

$$\lim_{\nu\to\infty} a_{jk,\nu} = 0, \quad 1 < k < j \le n,$$

and

$$\lim_{\nu\to\infty} a_{jj,\nu} = \lambda_j, \quad j = 1, \ldots, n.$$

Proof. **1.** Without loss of generality we may assume that $\|x_j\|_2 = 1$ for $j = 1, \ldots, n$. From Lemma 7.18 it follows that

$$\|P_{A^\nu S_m} - P_{T_m}\|_2 \le M r^\nu, \quad m = 1, \ldots, n-1, \quad \nu \in \mathbb{N}, \tag{7.16}$$

for some constant M and

$$r := \max_{m=1,\ldots,n-1} \left| \frac{\lambda_{m+1}}{\lambda_m} \right| < 1.$$

From this, for the projections

$$w_{m\nu} := P_{A^\nu S_m} x_m, \quad m = 1, \ldots, n-1,$$

and $w_{n\nu} := x_n$, we conclude that

$$\|w_{m\nu} - x_m\|_2 \le M r^\nu, \quad m = 1, \ldots, n, \quad \nu \in \mathbb{N}. \tag{7.17}$$

For sufficiently large ν the vectors $w_{1\nu}, \ldots, w_{n\nu}$ are linearly independent, and we have that

$$\operatorname{span}\{w_{1\nu}, \ldots, w_{m\nu}\} = A^\nu S_m, \quad m = 1, \ldots, n-1.$$

To prove this we assume to the contrary that the vectors $w_{1\nu}, \dots, w_{n\nu}$ are not linearly independent for all sufficiently large ν. Then there exists a sequence ν_ℓ such that the vectors $w_{1\nu_\ell}, \dots, w_{n\nu_\ell}$ are linearly dependent for each $\ell \in \mathbb{N}$. Hence there exist complex numbers $\alpha_{1\ell}, \dots, \alpha_{n\ell}$ such that

$$\sum_{k=1}^{n} \alpha_{k\ell} w_{kn_\ell} = 0 \quad \text{and} \quad \sum_{k=1}^{n} |\alpha_{k\ell}|^2 = 1. \tag{7.18}$$

By the Bolzano–Weierstrass theorem, without loss of generality, we may assume that

$$\alpha_{k\ell} \to \alpha_k, \quad \ell \to \infty, \quad k = 1, \dots, n.$$

Passing to the limit $\ell \to \infty$ in (7.18) with the aid of (7.17) now leads to

$$\sum_{k=1}^{n} \alpha_k x_k = 0 \quad \text{and} \quad \sum_{k=1}^{n} |\alpha_k|^2 = 1,$$

which contradicts the linear independence of the eigenvectors $x_1, \dots, x_n$.
2. We orthonormalize by setting $\tilde{p}_1 := x_1$ and

$$\tilde{p}_m := x_m - P_{T_{m-1}} x_m, \quad m = 2, \dots, n,$$

$$p_m := \frac{\tilde{p}_m}{\|\tilde{p}_m\|_2}, \quad m = 1, \dots, n,$$

and, analogously, $\tilde{v}_{1\nu} := w_{1\nu}$ and

$$\tilde{v}_{m\nu} := w_{m\nu} - P_{A^\nu S_{m-1}} w_{m\nu}, \quad m = 2, \dots, n,$$

$$v_{m\nu} := \frac{\tilde{v}_{m\nu}}{\|\tilde{v}_{m\nu}\|_2}, \quad m = 1, \dots, n.$$

Then

$$\operatorname{span}\{p_1, \dots, p_m\} = T_m, \quad m = 1, \dots, n-1,$$

and by repeating the above argument,

$$\operatorname{span}\{v_{1\nu}, \dots, v_{m\nu}\} = A^\nu S_m, \quad m = 1, \dots, n-1, \tag{7.19}$$

for sufficiently large ν. Writing

$$\tilde{p}_m - \tilde{v}_{m\nu} = x_m - w_{m\nu} + (P_{A^\nu S_{m-1}} - P_{T_{m-1}}) x_m + P_{A^\nu S_{m-1}} (w_{m\nu} - x_m),$$

with the aid of (7.16) and (7.17) we obtain that

$$\|\tilde{v}_{m\nu} - \tilde{p}_m\|_2 \le 3Mr^\nu, \quad m = 1, \dots, n, \quad \nu \in \mathbb{N}.$$

From this and the representation

$$v_{m\nu} - p_m = \frac{\tilde{v}_{m\nu}}{\|\tilde{v}_{m\nu}\|_2} \, \frac{\|\tilde{p}_m\|_2 - \|\tilde{v}_{m\nu}\|_2}{\|\tilde{p}_m\|_2} + \frac{\tilde{v}_{m\nu} - \tilde{p}_m}{\|\tilde{p}_m\|_2}$$

it follows that

$$\|v_{m\nu} - p_m\|_2 \le C r^\nu, \quad m = 1, \dots, n, \quad \nu \in \mathbb{N}, \tag{7.20}$$

for some constant C.

3. From (7.14) and (7.19), by induction, we deduce the existence of phase factors $\varphi_{m\nu} \in \mathbb{C}$ with $|\varphi_{m\nu}| = 1$ such that

$$q_{m\nu} = \varphi_{m\nu} v_{m\nu}, \quad m = 1, \dots, n.$$

Therefore, defining the diagonal matrices $D_\nu = \operatorname{diag}(\varphi_{1\nu}, \dots, \varphi_{n\nu})$ and the unitary matrices $V_\nu = (v_{1\nu}, \dots, v_{n\nu})$, we have the relation

$$D_\nu V_\nu = V_\nu D_\nu = \tilde{Q}_\nu.$$

This implies that

$$\begin{aligned} A_{\nu+1} &= \tilde{Q}_\nu^* A \tilde{Q}_\nu = D_\nu^* V_\nu^* A V_\nu D_\nu \\ &= D_\nu^*(V_\nu^* - P^*) A V_\nu D_\nu + D_\nu^* P^* A (V_\nu - P) D_\nu + D_\nu^* P^* A P D_\nu, \end{aligned} \tag{7.21}$$

where $P = (p_1, \dots, p_n)$. Because of (7.20) we have that

$$\|V_\nu - P\|_2 = \|V_\nu^* - P^*\|_2 \to 0, \quad \nu \to \infty.$$

Furthermore, $D_\nu^* P^* A P D_\nu$ is an upper triangular matrix with diagonal elements $\operatorname{diag}(\lambda_1, \dots, \lambda_n)$. Hence, the assertion of the theorem follows by passing to the limit $\nu \to \infty$ in (7.21). We note that for the elements above the diagonal we do not, in general, have convergence because of the occurrence of the phase factors. □

For the actual numerical implementation we have to describe the computation of $A_{\nu+1}$ according to (7.15). From page 20 we recall that orthonormalizing n vectors $a_1, \dots, a_n$ from left to right is equivalent to determining orthonormal vectors $q_1, \dots, q_n$ and an upper triangular matrix $R = (r_{jk})$ such that

$$a_k = \sum_{i=1}^{k} r_{ik} q_i, \quad k = 1, \dots, n.$$

For the matrices $A = (a_1, \dots, a_n)$ and $Q = (q_1, \dots, q_n)$ this corresponds to a QR decomposition

$$A = QR$$

as described in detail in Section 2.4. Now assume that $A_\nu = \tilde{Q}_{\nu-1}^* A \tilde{Q}_{\nu-1}$ has been determined according to (7.15). To generate $A_{\nu+1}$ from this, a QR decomposition of the matrix $A\tilde{Q}_{\nu-1}$ is required, since

$$A^\nu S_m = A A^{\nu-1} S_m = \operatorname{span}\{A q_{1,\nu-1}, \dots, A q_{m,\nu-1}\}.$$

This is obtained from a QR decomposition

$$A_\nu = Q_\nu R_\nu \tag{7.22}$$

of A_ν by

$$A\tilde{Q}_{\nu-1} = \tilde{Q}_{\nu-1} A_\nu = \tilde{Q}_{\nu-1} Q_\nu R_\nu = \tilde{Q}_\nu R_\nu,$$

where $\tilde{Q}_\nu = \tilde{Q}_{\nu-1} Q_\nu$. From this we find that

$$A_{\nu+1} = \tilde{Q}_\nu^* A \tilde{Q}_\nu = Q_\nu^* A_\nu Q_\nu = R_\nu Q_\nu. \tag{7.23}$$

Hence the two equations (7.22) and (7.23) represent one step of the successive iterations of subspaces as described in Theorem 7.19.

Now the QR algorithm consists in performing these iterations starting from the canonical basis $e_1, \ldots, e_n$, which means that in the first step a QR decomposition is required for $A_1 = A = (Ae_1, \ldots, Ae_n)$.

Theorem 7.20 *(QR algorithm) Let A be a diagonalizable matrix with eigenvalues*

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$$

and corresponding eigenvectors $x_1, x_2, \ldots, x_n$, and assume that

$$\operatorname{span}\{e_1, \ldots, e_m\} \cap \operatorname{span}\{x_{m+1}, \ldots, x_n\} = \{0\} \tag{7.24}$$

for $m = 1, \ldots, n-1$. Starting with $A_1 = A$, construct a sequence (A_ν) by determining a QR decomposition

$$A_\nu = Q_\nu R_\nu$$

and setting

$$A_{\nu+1} := R_\nu Q_\nu$$

for $\nu = 0, 1, 2, \ldots$. Then for $A_\nu = (a_{jk,\nu})$ we have convergence:

$$\lim_{\nu\to\infty} a_{jk,\nu} = 0, \quad 1 < k < j \le n,$$

and

$$\lim_{\nu\to\infty} a_{jj,\nu} = \lambda_j, \quad j = 1, \ldots, n.$$

Proof. This is just a special case of Theorem 7.19. □

We proceed with a discussion of the assumption (7.24). Define the matrices $X := (x_1, \ldots, x_n)$ and $Y := X^{-1} = (y_{jk})$. Then the identity $I = XY$ means that

$$e_j = \sum_{k=1}^n x_k y_{kj}, \quad j = 1, \ldots, n.$$

For fixed $m = 1, \dots, n-1$ the property (7.24) holds if and only if

$$\sum_{j=1}^{m} \alpha_j e_j \in \operatorname{span}\{x_{m+1}, \dots, x_n\}$$

implies that $\alpha_1 = \cdots = \alpha_m = 0$. This in turn is satisfied if and only if the homogeneous linear system

$$\sum_{j=1}^{m} y_{kj}\alpha_j = 0, \quad k = 1, \dots, m,$$

admits only the trivial solution, since

$$\sum_{j=1}^{k} \alpha_j e_j = \sum_{k=1}^{n} x_k \sum_{j=1}^{m} y_{kj}\alpha_j.$$

Hence (7.24) holds if and only if for $m = 1, \dots, n-1$ the $m \times m$ submatrices (y_{kj}), $k, j = 1, \dots, m$, are nonsingular. This means that for the matrix Y, Gaussian elimination works without interchanging columns; i.e., the matrix Y has an LR decomposition. Since Gaussian elimination with column pivoting always works, there exists a permutation matrix P such that we have an LR decomposition $PY = LR$ (see Problem 2.16). Hence it is plausible that the assumption (7.24) is not very restrictive. Indeed, it can be shown that convergence of the QR algorithm also holds when (7.24) is not satisfied. However, in general, the eigenvalues on the diagonal will not occur ordered according to their size (see [65]). Furthermore, it can be shown that in the case of eigenvalues with the same modulus, the QR algorithm still works in the sense of an appropriately modified version of Theorem 7.20. For example, for two conjugate complex eigenvalues, the upper rectangular matrix will be distorted through a two-by-two block on the diagonal. The blocks do not converge, but still the conjugate complex eigenvalues can be obtained as eigenvalues of the individual two-by-two blocks (see [65]).

In principle, the QR decomposition required in each step of the QR algorithm can be done through the Gram–Schmidt procedure. However, in practice, because of the ill-conditioning of the Gram–Schmidt procedure, orthogonalizing by Householder transformations is preferable. For details we refer back to Section 2.4.

The basic form of the QR algorithm as described above is not yet efficient enough for applications, since each iteration step requires $O(n^3)$ operations. The speed of convergence is determined by the location of the eigenvalues with respect to one another. The matrix $A - \sigma I$ has the eigenvalues $\lambda_j - \sigma$ for $j = 1, \dots, n$. If we choose for σ an approximate value of the eigenvalue λ_n of smallest absolute value, then $\lambda_n - \sigma$ becomes small. This will speed

up the convergence in the last row of the matrix, since

$$\frac{|\lambda_n - \sigma|}{|\lambda_{n-1} - \sigma|} \ll 1.$$

Having reduced the elements of the last row to almost zero, the last row and column of the matrix may be neglected. This means that the smallest eigenvalue is deflated by canceling the last row and column, and the same procedure can be applied to the remaining $(n-1) \times (n-1)$ matrix with the parameter σ changed to be close to λ_{n-1}. This so-called *shift* and *deflation strategy* leads to a tremendous speeding up of the convergence. For details we refer to [27, 65].

The computational costs of one step of the QR algorithm is reduced when the matrix has a large number of zero entries. For example, for tridiagonal matrices all matrices generated in the QR algorithm remain tridiagonal. In the following section we will consider so-called Hessenberg matrices, which differ from upper triangular matrices only by a non-zero first subdiagonal. It can be shown (see Problem 7.16) that the Hessenberg form is also invariant with respect to the QR algorithm. Hence, for practical computations it is convenient first to transform the matrix into Hessenberg form.

In general, comparing the computational costs, for symmetric matrices the QR algorithm is superior to the Jacobi method. However, the actual programming for the Jacobi method is very simple as compared with the QR algorithm. Hence for small matrix size n the Jacobi method is still attractive.

7.5 Hessenberg Matrices

Definition 7.21 *An $n \times n$ matrix $B = (b_{jk})$ is called a* Hessenberg matrix *if $b_{jk} = 0$ for $1 \leq k \leq j-2$, $j = 3, \ldots, n$; i.e., in the lower triangular part of a Hessenberg matrix only the elements of the first subdiagonal can be different from zero.*

We proceed by showing that each matrix A can be transformed into Hessenberg form by unitary transformations using Householder matrices. We start with generating zeros in the first column by multiplying A from the left by a Householder matrix H_1. We write

$$A = \begin{pmatrix} a_{11} & * \\ \tilde{a}_1 & \tilde{A} \end{pmatrix},$$

where $\tilde{A}$ is an $(n-1) \times (n-1)$ matrix and $\tilde{a}_1$ an $(n-1)$ vector. Then considering a Householder matrix H_1 of the form

$$H_1 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{H}_1 \end{pmatrix},$$

where $\tilde{H}_1 = I - 2\tilde{v}_1\tilde{v}_1^*$ is an $(n-1)\times(n-1)$ Householder matrix, we have

$$AH_1^* = \begin{pmatrix} a_{11} & * \\ \tilde{a}_1 & \tilde{A}\tilde{H}_1^* \end{pmatrix}$$

and

$$H_1AH_1^* = \begin{pmatrix} a_{11} & * \\ \tilde{H}_1\tilde{a}_1 & \tilde{H}_1\tilde{A}\tilde{H}_1^* \end{pmatrix}.$$

As shown in the proof of Theorem 2.13, choosing

$$\tilde{v}_1 = \frac{u_1}{\sqrt{u_1^*u_1}},$$

where

$$u_1 = \tilde{a}_1 \mp \sigma\,(1,0,\dots,0)^T$$

and

$$\sigma = \begin{cases} \dfrac{a_{21}}{|a_{21}|}\sqrt{\tilde{a}_1^*\tilde{a}_1}\,, & a_{21} \neq 0, \\ \sqrt{\tilde{a}_1^*\tilde{a}_1}\,, & a_{21} = 0, \end{cases}$$

eliminates all elements of $\tilde{a}_1$ with the exception of the first component. Hence the first column of the transformed matrix is of the required form.

Now assume that A_k is an $n \times n$ matrix of the form

$$A_k = \begin{pmatrix} B_k & * \\ 0\ \tilde{a}_k & \tilde{A}_{n-k} \end{pmatrix},$$

where B_k is a $k \times k$ Hessenberg matrix, $\tilde{A}_{n-k}$ an $(n-k)\times(n-k)$ matrix, $\tilde{a}_k$ an $(n-k)$ vector, and 0 the $(n-k)\times(k-1)$ zero matrix. Then for a Householder transformation of the form

$$H_k = \begin{pmatrix} I_k & 0 \\ 0 & \tilde{H}_{n-k} \end{pmatrix},$$

where I_k denotes the $k\times k$ identity matrix and $\tilde{H}_{n-k}$ is an $(n-k)\times(n-k)$ Householder matrix, it follows that

$$A_kH_k^* = \begin{pmatrix} B_k & * \\ 0\ \tilde{a}_k & \tilde{A}_{n-k}\tilde{H}_{n-k}^* \end{pmatrix}$$

and

$$H_kA_kH_k^* = \begin{pmatrix} B_k & * \\ 0\ \tilde{H}_{n-k}\tilde{a}_k & \tilde{H}_{n-k}\tilde{A}_{n-k}\tilde{H}_{n-k}^* \end{pmatrix}.$$

Now, proceeding as above, we can choose $\tilde{H}_{n-k}$ such that all elements of $\tilde{H}_{n-k}\tilde{a}_k$ vanish with the exception of the first component. This procedure reduces a further column into Hessenberg form. We can summarize our analysis in the following theorem.

Theorem 7.22 *To each $n \times n$ matrix A there exist $n-2$ Householder matrices $H_1, \ldots, H_{n-2}$ such for $Q = H_{n-2} \cdots H_1$ the matrix*

$$B = Q^* A Q$$

is a Hessenberg matrix.

For a Hessenberg matrix the value of the characteristic polynomial and its derivative at a point $\lambda \in \mathbb{C}$ can be computed easily without computing the coefficients of the polynomial. These two quantities are required for employing Newton's method for approximating the eigenvalues as the zeros of the characteristic polynomial. We first consider the case of a symmetric Hessenberg matrix.

Example 7.23 *Let*

$$A = \begin{pmatrix} a_1 & c_2 & & & & \\ c_2 & a_2 & c_3 & & & \\ & c_3 & a_3 & c_4 & & \\ & & \cdot & \cdot & \cdot & \\ & & & c_{n-1} & a_{n-1} & c_n \\ & & & & c_n & a_n \end{pmatrix}$$

be a symmetric tridiagonal matrix. Denote by A_k the $k \times k$ submatrix consisting of the first k rows and columns of A, and let p_k denote the characteristic polynomial of A_k. Then we have the recurrence relations

$$p_k(\lambda) = (a_k - \lambda) p_{k-1}(\lambda) - c_k^2 p_{k-2}(\lambda), \quad k = 2, \ldots, n, \tag{7.25}$$

and

$$p_k'(\lambda) = (a_k - \lambda) p_{k-1}'(\lambda) - c_k^2 p_{k-2}'(\lambda) - p_{k-1}(\lambda), \quad k = 2, \ldots, n, \tag{7.26}$$

starting with $p_0(\lambda) = 1$ and $p_1(\lambda) = a_1 - \lambda$.

Proof. The recursion (7.25) follows by expanding $\det(A_k - \lambda I)$ with respect to the last column, and (7.26) is obtained by differentiating (7.25). □

Example 7.24 The $n \times n$ tridiagonal matrix

$$A = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}$$

has the eigenvalues

$$\lambda_j = 4 \sin^2 \frac{j\pi}{2(n+1)}, \quad j = 1, \ldots, n$$

(see Example 4.17). Table 7.1 gives the results of the Newton iteration using (7.25) and (7.26) for computing the smallest eigenvalue $\lambda_{\min} = \lambda_1$ and the largest eigenvalue $\lambda_{\max} = \lambda_n$ for $n = 10$. The starting values are obtained from the Gerschgorin estimates $|\lambda - 2| \leq 2$ following from Theorem 7.7. □

TABLE 7.1. Hessenberg method for Example 7.24

$\lambda_{\max}$	$\lambda_{\min}$
4.00000000	0.00000000
3.95000000	0.05000000
3.92542110	0.07457890
3.91933549	0.08066451
3.91898705	0.08101295
3.91898595	0.08101405
3.91898595	0.08101405

We conclude this section by describing the computation of the quotient of the value of the characteristic polynomial $p(\lambda) = \det(B - \lambda I)$ and its derivative for a general Hessenberg matrix $B = (b_{jk})$. We assume that $b_{j,j-1} \neq 0$ for $j = 2, \dots, n$; i.e., B is irreducible (see Problem 7.15). For a given λ we determine

$$\xi = \xi(\lambda) = (\xi_1, \dots, \xi_n)^T$$

and $\alpha = \alpha(\lambda)$ such that

$$\begin{array}{rcccl}
(b_{11} - \lambda)\xi_1 + & b_{12}\xi_2 + & \cdots & + & b_{1n}\xi_n = \alpha, \\
b_{21}\xi_1 + & (b_{22} - \lambda)\xi_2 + & \cdots & + & b_{2n}\xi_n = 0, \\
 & & \cdots & & \\
 & & b_{n,n-1}\xi_{n-1} & + & (b_{nn} - \lambda)\xi_n = 0,
\end{array}$$

and $\xi_n = 1$. This is an $n \times n$ upper triangular linear system for the n unknowns $\alpha, \xi_1, \dots, \xi_{n-1}$, and it can be solved by backward substitution.

Setting

$$C = \begin{pmatrix} b_{11} - \lambda & b_{12} & \cdot & \cdot & b_{1,n-1} & \alpha \\ b_{21} & b_{21} - \lambda & \cdot & \cdot & b_{2,n-1} & 0 \\ & & \cdot & \cdot & \cdot & \cdot \\ & & & & b_{n,n-1} & 0 \end{pmatrix},$$

by Cramer's rule we have that

$$1 = \xi_n = \frac{\det C}{\det(B - \lambda I)} = \frac{(-1)^{n-1} b_{21} \cdots b_{n,n-1} \alpha}{\det(B - \lambda I)},$$

that is,

$$p(\lambda) = (-1)^{n-1} b_{21} \cdots b_{n,n-1} \alpha(\lambda).$$

Differentiating the last equation yields

$$p'(\lambda) = (-1)^{n-1} b_{21} \cdots b_{n,n-1} \alpha'(\lambda),$$

and therefore

$$\frac{p(\lambda)}{p'(\lambda)} = \frac{\alpha(\lambda)}{\alpha'(\lambda)} .$$

By differentiating the above linear system with respect to λ we obtain the linear system

$$\begin{aligned}
(b_{11} - \lambda)\eta_1 + \quad b_{12}\eta_2 + \cdots + b_{1,n-1}\eta_{n-1} &= \xi_1 + \beta, \\
b_{21}\eta_1 + (b_{22} - \lambda)\eta_2 + \cdots + b_{2,n-1}\eta_{n-1} &= \xi_2, \\
\cdots \\
b_{n,n-1}\eta_{n-1} &= \xi_n
\end{aligned}$$

for the derivatives $\beta = \alpha', \eta_1 = \xi_1', \ldots, \eta_{n-1} = \xi_{n-1}'$. This linear system again can be solved by backward substitution for the n unknowns $\beta, \eta_1, \ldots, \eta_{n-1}$. Thus we have proven the following theorem.

Theorem 7.25 *Let $B = (b_{jk})$ be an irreducible Hessenberg matrix and let $\lambda \in \mathbb{C}$. Starting from $\xi_n = 1$, $\eta_n = 0$, compute recursively*

$$\xi_{n-k} = \frac{1}{b_{n-k+1,n-k}} \left\{ \lambda \xi_{n-k+1} - \sum_{j=n-k+1}^{n} b_{n-k+1,j} \xi_j \right\},$$

$$\eta_{n-k} = \frac{1}{b_{n-k+1,n-k}} \left\{ \xi_{n-k+1} + \lambda \eta_{n-k+1} - \sum_{j=n-k+1}^{n} b_{n-k+1,j} \eta_j \right\}$$

for $k = 1, \ldots, n-1$ and

$$\alpha = -\lambda \xi_1 + \sum_{j=1}^{n} b_{1j} \xi_j,$$

$$\beta = -\xi_1 - \lambda \eta_1 + \sum_{j=1}^{n} b_{1j} \eta_j .$$

Then for the characteristic polynomial of B we have

$$\frac{p(\lambda)}{p'(\lambda)} = \frac{\alpha}{\beta} .$$

Problems

7.1 For the eigenvalues (repeated according to their algebraic multiplicity) of an $n \times n$ matrix A show that

$$\operatorname{tr} A = \sum_{j=1}^{n} \lambda_j \quad \text{and} \quad \det A = \prod_{j=1}^{n} \lambda_j .$$

7.2 For Example 7.1 show that in the case $p = 0$ the eigenvalues of the matrix A converge to the eigenvalues of the differential operator D as $n \to \infty$.

7.3 Show that the eigenvalues of the adjoint matrix A^* are the complex conjugate of the eigenvalues of the matrix A.

7.4 Show that for the eigenvalues of Hermitian matrices the geometric and the algebraic multiplicities coincide.

7.5 Use Gerschgorin's Theorem 7.7 to determine the approximate location of the eigenvalues of the matrix

$$A = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 5 & 1 \\ -2 & -1 & 9 \end{pmatrix} .$$

To check the estimates, compute the eigenvalues by finding the zeros of the characteristic polynomial.

7.6 Let A be a diagonalizable $n \times n$ matrix with eigenvalues $\lambda_1, \dots, \lambda_n$, B an $n \times n$ matrix, and λ an eigenvalue of $A + B$. Show that

$$\min_{j=1,\dots,n} |\lambda - \lambda_j| \le \|C\|_p \|C^{-1}\|_p \|B\|_p ,$$

where C is a nonsingular matrix such that $C^{-1}AC$ is diagonal and $p = 1, 2, \infty$.

7.7 Show that the Frobenius norm is indeed a norm on the linear space of matrices.

7.8 Write a computer program for the Jacobi method and test it for various examples.

7.9 Assume that A is a real symmetric $n \times n$ matrix with eigenvalue λ of multiplicity $n-1$ and a further eigenvalue $\mu \neq \lambda$. Show that $A = \lambda I + (\mu - \lambda) x x^*$, where $x^* x = 1$ and that by at most $n - 1$ Jacobi transformations A becomes diagonal.

7.10 Show convergence of the cyclic Jacobi method with threshold $[N(A)]^2/(2n^2)$.

7.11 Let A be a diagonalizable $n \times n$ matrix with eigenvalues $\lambda_1, \dots, \lambda_n$ and eigenvectors $x_1, \dots, x_n$, and assume that $|\lambda_1| > |\lambda_2| \ge |\lambda_3| \ge \cdots \ge |\lambda_n|$. Starting from $v_0 \in \mathbb{C}^n$ with $v_0 \notin \operatorname{span}\{x_2, \dots, x_n\}$ show that the sequence

$$v_{\nu+1} := \frac{A v_\nu}{\|A v_\nu\|_2} , \quad \nu = 0, 1, 2, \dots ,$$

is well-defined and that the sequence of Rayleigh quotients

$$R_\nu := \frac{(Av_\nu, v_\nu)}{\|v_\nu\|_2^2}, \quad \nu = 0, 1, 2, \dots,$$

satisfies the estimate

$$|R_\nu - \lambda_1| \leq C\, r^\nu, \quad \nu = 0, 1, 2, \dots,$$

for some constant $C > 0$ and $r := |\lambda_2/\lambda_1|$.

7.12 The matrix

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 1.5 & 1 & 1.5 \\ 2 & 0.5 & 0.5 \end{pmatrix}$$

has eigenvalue $\lambda = 4$ with eigenvector $x = (1,1,1)^T$. Construct a Householder matrix H such that

$$HAH^* = \begin{pmatrix} 4 & * & * \\ 0 & * & * \\ 0 & * & * \end{pmatrix}$$

and determine the remaining eigenvalues.

7.13 Write a computer program for the QR algorithm and test it for various examples.

7.14 Verify the numerical results of Table 5.2 for the Hilbert matrix.

7.15 Show that Hessenberg matrices $B = (b_{jk})$ with $b_{j,j-1} \neq 0$ for $j = 2, \dots, n$ are irreducible.

7.16 Show that the Hessenberg form of a matrix is preserved by the QR algorithm.

7.17 Show that the number of multiplications required for the transformation of a matrix into Hessenberg form via Householder transformations according to Theorem 7.22 is $5n^3/3 + O(n^2)$.

7.18 Write a computer program for transforming a matrix into Hessenberg form via Householder transformations according to Theorem 7.22.

7.19 Discuss Newton's method for the solution of $Ax = \lambda x$, $x^T x = 1$ in the neighborhood of a simple eigenvalue of a real symmetric matrix A.

7.20 Prove the inequality

$$\sum_{j=1}^n |\lambda_j|^2 \leq \left(\|A\|_F^4 - \frac{1}{2}\, \|AA^* - A^*A\|_F^2 \right)^{1/2}$$

for the eigenvalues of an $n \times n$ matrix A (see Corollary 7.9 and [41]).

8
Interpolation

Polynomials have attracted the attention of mathematicians for centuries because of their many beautiful properties. For numerical purposes they have the advantage that their computation reduces to additions and multiplications only. Therefore, it is quite natural to use polynomials for the approximation of more complicated functions. A classical approach to specifying the coefficients of a polynomial of degree n is to prescribe that its values at $n + 1$ distinct points coincide with those of the function to be approximated. The development and investigation of such interpolation polynomials has a long mathematical history, beginning with the use of the method of interpolation to tabulate the logarithms, as proposed by Briggs in the early seventeenth century.

It is the purpose of the first section, Section 8.1, of this chapter to introduce the classical theory of polynomial interpolation, including discussions on the effective numerical computation of interpolation polynomials and an analysis of the resulting approximation error. The next section, Section 8.2, describes the corresponding theory for the interpolation of periodic functions by trigonometric polynomials. For a detailed study of the foundations of classical interpolation theory we refer to [16].

In the last two sections, Sections 8.3 and 8.4, we proceed with a study of interpolation by splines, i.e., piecewise polynomial interpolation, which was developed within the last fifty years and has turned into a successful tool in approximation theory and other parts of numerical analysis. For a comprehensive study of spline functions we refer to [18, 53], and for their use in computer-aided geometric design we refer to [23].

We would like to point out that interpolation is not only important as a tool for the approximation of functions that are difficult to compute or whose values are known only at discrete points. It also serves as an essential ingredient for developing numerical integration rules and methods for the approximate solution of differential and integral equations, as we shall see in the following chapters.

8.1 Polynomial Interpolation

For $n \in \mathbb{N} \cup \{0\}$, we denote by P_n the linear space of polynomials

$$p(x) = \sum_{k=0}^{n} a_k x^k$$

for a real (or complex) variable x and with real (or complex) coefficients $a_0, \dots, a_n$. A polynomial $p \in P_n$ is said to be of degree n if $a_n \neq 0$. In this chapter, we consider P_n as a subspace of the linear space $C[a, b]$ of continuous real- (or complex-) valued functions on the interval $[a, b]$, where $a < b$. For $m \in \mathbb{N}$ we denote by $C^m[a, b]$ the linear space of m times continuously differentiable real- (or complex-) valued functions on $[a, b]$.

We recall the following basic uniqueness property of algebraic polynomials as part of the fundamental theorem of algebra. Since we will use this property frequently, it is appropriate to include a simple proof by induction.

Theorem 8.1 *For $n \in \mathbb{N} \cup \{0\}$, each polynomial in P_n that has more than n (complex) zeros, where each zero is counted repeatedly according to its multiplicity, must vanish identically; i.e., all its coefficients must be equal to zero.*

Proof. Obviously, the statement is true for $n = 0$. Assume that it has been proven for some $n \geq 0$. By using the binomial formula for $x^k = [(x - z) + z]^k$ we can rewrite the polynomial $p \in P_{n+1}$ in the form

$$p(x) = \sum_{k=1}^{n+1} b_k (x - z)^k + b_0$$

with the coefficients $b_0, b_1, \dots, b_{n+1}$ depending on $a_0, a_1, \dots, a_{n+1}$ and z. If z is a zero of p, then we must have $b_0 = 0$, and this implies that $p(x) = (x - z)q(x)$ with $q \in P_n$. Obviously, q has more than n zeros, since p has more than $n + 1$ zeros. Hence, by the induction assumption, q must vanish identically, and this implies that p vanishes identically. □

Theorem 8.2 *The monomials $u_k(x) := x^k$, $k = 0, \dots, n$, are linearly independent.*

Proof. In order to prove this, assume that

$$\sum_{k=0}^{n} a_k u_k = 0,$$

that is,

$$\sum_{k=0}^{n} a_k x^k = 0, \quad x \in [a, b].$$

Then the polynomial with coefficients $a_0, a_1, \ldots, a_n$ has more than n distinct zeros, and from Theorem 8.1 it follows that all the coefficients must be zero. □

The linear independence of the monomials $u_0, \ldots, u_n$ implies that they form a basis for P_n and that P_n has dimension $n+1$.

Theorem 8.3 *Given $n+1$ distinct points $x_0, \ldots, x_n \in [a, b]$ and $n+1$ values $y_0, \ldots, y_n \in \mathbb{R}$, there exists a unique polynomial $p_n \in P_n$ with the property*

$$p_n(x_j) = y_j, \quad j = 0, \ldots, n. \tag{8.1}$$

In the Lagrange representation, *this* interpolation polynomial *is given by*

$$p_n = \sum_{k=0}^{n} y_k \ell_k \tag{8.2}$$

with the Lagrange factors

$$\ell_k(x) = \prod_{\substack{i=0 \\ i \neq k}}^{n} \frac{x - x_i}{x_k - x_i}, \quad k = 0, \ldots, n.$$

Proof. We note that $\ell_k \in P_n$ for $k = 0, \ldots, n$ and that the equations

$$\ell_k(x_j) = \delta_{jk}, \quad j, k = 0, \ldots, n, \tag{8.3}$$

hold, where $\delta_{jk} = 1$ for $k = j$, and $\delta_{jk} = 0$ for $k \neq j$. It follows that p_n given by (8.2) is in P_n, and it fulfills the required interpolation conditions $p_n(x_j) = y_j$, $j = 0, \ldots, n$.

To prove uniqueness of the interpolation polynomial we assume that $p_{n,1}, p_{n,2} \in P_n$ are two polynomials satisfying (8.1). Then the difference $p_n := p_{n,1} - p_{n,2}$ satisfies $p_n(x_j) = 0$, $j = 0, \ldots, n$; i.e., the polynomial $p_n \in P_n$ has $n+1$ zeros and therefore by Theorem 8.1 must be identically zero. This implies that $p_{n,1} = p_{n,2}$. □

The representation (8.2), which was discovered by Lagrange in 1794, is very convenient for theoretical investigations because of its simple structure. However, for practical computations it is suitable only for small n. For

n large the Lagrange factors become very large and highly oscillatory, which causes ill-conditioning of the Lagrange interpolation polynomial. Already in 1676, in his study of quadrature formulae (see Theorem 9.3), Newton had obtained a representation of the interpolation polynomial that is more practical for computational purposes. For its description we need to give the following definition.

Definition 8.4 *Given $n+1$ distinct points $x_0, \dots, x_n \in [a,b]$ and $n+1$ values $y_0, \dots, y_n \in \mathbb{R}$, the* divided differences *D_j^k of order k at the point x_j are recursively defined by*

$$D_j^0 := y_j, \quad j = 0, \dots, n,$$

$$D_j^k := \frac{D_{j+1}^{k-1} - D_j^{k-1}}{x_{j+k} - x_j}, \quad j = 0, \dots, n-k, \quad k = 1, \dots, n.$$

We notice that the points $x_0, \dots, x_n$ need not be in ascending order. It is convenient to arrange the divided differences according to the tableau

$$\begin{array}{llllll}
x_0 & y_0 = D_0^0 & & & \\
 & & D_0^1 & & \\
x_1 & y_1 = D_1^0 & & D_0^2 & \\
 & & D_1^1 & & D_0^3 \\
x_2 & y_2 = D_2^0 & & D_1^2 & \\
 & & D_2^1 & & \\
x_3 & y_3 = D_3^0 & & &
\end{array}$$

which we illustrate by the following example. Obviously, for the full tableau the computational cost is of order $O(n^2)$.

Example 8.5 For the points $x_0 = 0$, $x_1 = 1$, $x_2 = 3$, $x_4 = 4$ and the values $y_0 = 0$, $y_1 = 2$, $y_2 = 8$, $y_4 = 9$ the tableau of the divided differences is given by

$$\begin{array}{cccccc}
0 & 0 & & & \\
 & & 2 & & \\
1 & 2 & & 1/3 & \\
 & & 3 & & -1/4 \\
3 & 8 & & -2/3 & \\
 & & 1 & & \\
4 & 9 & & &
\end{array}$$

Each value D_j^k in the kth column is obtained by taking the difference of the two neighboring values D_{j+1}^{k-1} and D_j^{k-1} in the preceding column and dividing it by the difference $x_{j+k} - x_j$ of the points x_{j+k} and x_j. □

Lemma 8.6 *The divided differences satisfy the relation*

$$D_j^k = \sum_{m=j}^{j+k} y_m \prod_{\substack{i=j \\ i\neq m}}^{j+k} \frac{1}{x_m - x_i}, \quad j = 0, \dots, n-k, \quad k = 1, \dots, n. \tag{8.4}$$

Proof. We proceed by induction with respect to the order k. Trivially, (8.4) holds for $k = 1$. We assume that (8.4) has been proven for order $k-1$ for some $k \geq 2$. Then, using Definition 8.4, the induction assumption, and the identity

$$\frac{1}{x_{j+k} - x_j} \left\{ \frac{1}{x_m - x_{j+k}} - \frac{1}{x_m - x_j} \right\} = \frac{1}{(x_m - x_{j+k})(x_m - x_j)},$$

we obtain

$$D_j^k = \frac{1}{x_{j+k} - x_j} \left\{ \sum_{m=j+1}^{j+k} y_m \prod_{\substack{i=j+1 \\ i\neq m}}^{j+k} \frac{1}{x_m - x_i} - \sum_{m=j}^{j+k-1} y_m \prod_{\substack{i=j \\ i\neq m}}^{j+k-1} \frac{1}{x_m - x_i} \right\}$$

$$= \frac{1}{x_{j+k} - x_j} \sum_{m=j+1}^{j+k-1} y_m \left\{ \frac{1}{x_m - x_{j+k}} - \frac{1}{x_m - x_j} \right\} \prod_{\substack{i=j+1 \\ i\neq m}}^{j+k-1} \frac{1}{x_m - x_i}$$

$$+ y_{j+k} \prod_{i=j}^{j+k-1} \frac{1}{x_{j+k} - x_i} + y_j \prod_{i=j+1}^{j+k} \frac{1}{x_j - x_i} = \sum_{m=j}^{j+k} y_m \prod_{\substack{i=j \\ i\neq m}}^{j+k} \frac{1}{x_m - x_i};$$

i.e., (8.4) also holds for order k. □

Theorem 8.7 *In the* Newton representation, *for $n \geq 1$ the uniquely determined interpolation polynomial p_n of Theorem 8.3 is given by*

$$p_n(x) = y_0 + \sum_{k=1}^{n} D_0^k \prod_{i=0}^{k-1} (x - x_i). \tag{8.5}$$

Proof. We denote the right-hand side of (8.5) by $\tilde{p}_n$ and establish $p_n = \tilde{p}_n$ by induction with respect to the degree n. For $n = 1$ the representation (8.5) is correct. We assume that (8.5) has been proven for degree $n-1$ for some $n \geq 2$ and consider the difference $d_n := p_n - \tilde{p}_n$. Since

$$d_n(x) = p_n(x) - \tilde{p}_{n-1}(x) - D_0^n \prod_{i=0}^{n-1} (x - x_i),$$

as a consequence of Theorem 8.3 and Lemma 8.6 the coefficient of x^n in the polynomial d_n vanishes; i.e., $d_n \in P_{n-1}$. Using the induction assumption, we have that

$$\tilde{p}_{n-1}(x_j) = y_j = p_n(x_j), \quad j = 0, \dots, n-1,$$

and therefore

$$d_n(x_j) = 0, \quad j = 0, \dots, n-1.$$

Hence, by Theorem 8.1 it follows that $d_n = 0$, and therefore $p_n = \tilde{p}_n$. □

Example 8.8 The interpolation polynomial corresponding to Example 8.5 is given by

$$p_3(x) = 2x + \frac{1}{3}x(x-1) - \frac{1}{4}x(x-1)(x-3).$$ □

Analogously to the Horner scheme (see (6.10)), the value of the Newton interpolation polynomial at a point x can be obtained by nested multiplications according to

$$\begin{aligned} p_n(x) &= a_n(x-x_0)(x-x_1)\cdots(x-x_{n-1}) + \cdots + a_1(x-x_0) + a_0 \\ &= (\dots(a_n(x-x_{n-1}) + a_{n-1})(x-x_{n-2}) + \cdots + a_1)(x-x_0) + a_0 \end{aligned}$$

by $O(n)$ multiplications and additions. For an evaluation of the interpolation polynomial at a single point x without explicitly computing the coefficients of the polynomial, the following *Neville scheme* is very practical. From the formal coincidence of the recursion (8.6) and Definition 8.4 for the divided differences, it is obvious that the computations for (8.6) can be arranged in a tableau analogous to the tableau for the divided differences.

Theorem 8.9 *Given $n+1$ distinct points $x_0, \dots, x_n \in [a,b]$ and $n+1$ values $y_0, \dots, y_n \in \mathbb{R}$, the uniquely determined interpolation polynomials $p_i^k \in P_k$, $i = 0, \dots, n-k$, $k = 0, \dots, n$, with the interpolation property*

$$p_i^k(x_j) = y_j, \quad j = i, \dots, i+k,$$

satisfy the recursive relation

$$\begin{aligned} p_i^0(x) &= y_i, \\ p_i^k(x) &= \frac{(x-x_i)p_{i+1}^{k-1}(x) - (x-x_{i+k})p_i^{k-1}(x)}{x_{i+k}-x_i}, \quad k = 1, \dots, n. \end{aligned} \tag{8.6}$$

Proof. We again proceed by induction with respect to the degree k. Obviously, the statement is true for $k = 1$. Assume that the assertion has been

proven for degree $k-1$ for some $k \geq 2$. Then the right-hand side of (8.6) describes a polynomial $p \in P_k$, and by the induction assumption we find that the interpolation conditions

$$p(x_j) = \frac{(x_j - x_i)y_j - (x_j - x_{i+k})y_j}{x_{i+k} - x_i} = y_j, \quad j = i+1, \dots, i+k-1,$$

as well as $p(x_i) = y_i$ and $p(x_{i+k}) = y_{i+k}$ are fulfilled. □

The main application of polynomial interpolation consists in the approximation of continuous functions $f : [a,b] \to \mathbb{R}$. In this case, given $n+1$ distinct points $x_0, \dots, x_n \in [a,b]$, by

$$L_n : C[a,b] \to P_n$$

we denote the *interpolation operator* that maps the function $f \in C[a,b]$ onto its uniquely determined interpolation polynomial $L_n f \in P_n$ with the property

$$(L_n f)(x_j) = f(x_j), \quad j = 0, \dots, n. \tag{8.7}$$

From the Lagrange representation (8.2) it can be seen that the operator L_n is linear and bounded (see Problem 8.4). Moreover, since $L_n p = p$ for all $p \in P_n$, the interpolation operator is a projection; i.e., $L_n^2 = L_n$.

The interpolation polynomial $L_n f$ is used as an approximation for the function f, since in general, the polynomial $L_n f$ is better suited for computational purposes than the original function f. In the sequel we shall be concerned with estimating the approximation error $f - L_n f$.

Theorem 8.10 *Let $f : [a,b] \to \mathbb{R}$ be $(n+1)$-times continuously differentiable. Then the* remainder *$R_n f := f - L_n f$ for polynomial interpolation with $n+1$ distinct points $x_0, \dots, x_n \in [a,b]$ can be represented in the form*

$$(R_n f)(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{j=0}^{n} (x - x_j), \quad x \in [a,b], \tag{8.8}$$

for some $\xi \in [a,b]$ depending on x.

Proof. Since (8.8) is trivially satisfied if x coincides with one of the interpolation points $x_0, \dots, x_n$, we need be concerned only with the case where x does not coincide with one of the interpolation points. We define

$$q_{n+1}(x) := \prod_{j=0}^{n} (x - x_j)$$

and, keeping x fixed, consider $g : [a,b] \to \mathbb{R}$ given by

$$g(y) := f(y) - (L_n f)(y) - q_{n+1}(y)\, \frac{f(x) - (L_n f)(x)}{q_{n+1}(x)}, \quad y \in [a,b].$$

By the assumption on f, the function g is also $(n+1)$-times continuously differentiable. Obviously, g has at least $n+2$ zeros, namely x and $x_0, \dots, x_n$. Then, by Rolle's theorem the derivative g' has at least $n+1$ zeros. Repeating the argument, by induction we deduce that the derivative $g^{(n+1)}$ has at least one zero in $[a, b]$, which we denote by ξ. For this zero we have that

$$0 = f^{(n+1)}(\xi) - (n+1)! \, \frac{(R_n f)(x)}{q_{n+1}(x)} \, ,$$

and from this we obtain (8.8). □

The intermediate point ξ in the error representation (8.8) is not, in general, known explicitly. Therefore, the interpolation error is estimated by the following corollary.

Corollary 8.11 *Under the assumptions of Theorem 8.10 we have the error estimate*

$$\|R_n f\|_\infty \leq \frac{1}{(n+1)!} \, \|q_{n+1}\|_\infty \, \|f^{(n+1)}\|_\infty .$$

Example 8.12 The *linear interpolation* is given by

$$(L_1 f)(x) = \frac{1}{h} \, [f(x_0)(x_1 - x) + f(x_1)(x - x_0)]$$

with the step width $h = x_1 - x_0$. For the polynomial $q_2(x) = (x - x_0)(x - x_1)$ we have that

$$\max_{x \in [x_0, x_1]} |q_2(x)| = \frac{h^2}{4} \, .$$

Therefore, by Corollary 8.11, the error occurring in linear interpolation of a twice continuously differentiable function f can be estimated by

$$|(R_1 f)(x)| \leq \frac{h^2}{8} \max_{y \in [x_0, x_1]} |f''(y)|, \quad x \in [x_0, x_1]. \tag{8.9}$$

For example, the error in linear interpolation with step size $h = 0.01$ for the sine function is less than or equal to $h^2/8 = 0.0000125$. □

By the following examples we want to introduce the question of whether the interpolation polynomials converge when the number $n+1$ of interpolation points, and hence the degree n of the interpolation polynomials, tends to infinity.

Example 8.13 Let $f(x) := \sin x$ and let $x_0, \dots, x_n \in [0, \pi]$ be $n+1$ distinct points. Since

$$|f^{(n+1)}(x)| \leq 1, \quad x \in [0, \pi],$$

and

$$|q_{n+1}(x)| \leq \pi^{n+1}, \quad x \in [0, \pi],$$

by Corollary 8.11, we have the estimate

$$|(R_n f)(x)| \le \frac{\pi^{n+1}}{(n+1)!}, \quad x \in [0, \pi].$$

Hence the sequence $(L_n f)$ of interpolation polynomials converges to the interpolated function f uniformly on $[0, \pi]$ as $n \to \infty$. □

Example 8.14 A first detailed example of the insufficiency of polynomial interpolation even for analytic functions was investigated by Runge in 1901. He considered the simple function

$$f(x) = \frac{1}{1 + 25x^2}$$

on the interval $[-1, 1]$ with equidistant interpolation points. He discovered that as the degree n tends to infinity, the interpolation polynomials diverge for $0.726 \le |x| \le 1$, whereas the approximation works satisfactorily in the central portion of the interval (see Problem 8.6). Although f is analytic in all of $\mathbb{R}$, its poles in the complex plane at $\pm i/5$ are responsible for this divergence. □

Example 8.15 Consider the continuous function

$$f(x) := \begin{cases} x \sin \dfrac{\pi}{x}, & x \in (0, 1], \\ 0, & x = 0. \end{cases}$$

With the interpolation points chosen as

$$x_j = \frac{1}{j+1}, \quad j = 0, \dots, n,$$

we have that $f(x_j) = 0$, $j = 0, \dots, n$, and therefore $L_n f = 0$ for all n. Hence, in this case the sequence $(L_n f)$ converges only at the points x_j, $j \in \mathbb{N} \cup \{0\}$, to the interpolated function f. □

These three examples illustrate that for polynomial interpolation both convergence and divergence are possible. We complement the examples by stating the following two theorems without detailed proofs.

Theorem 8.16 (Marcinkiewicz) *For each function $f \in C[a, b]$ there exists a sequence of interpolation points $(x_j^{(n)})$, $j = 0, \dots, n$, $n = 0, 1, \dots$, such that the sequence $(L_n f)$ of interpolation polynomials $L_n f \in P_n$ with $(L_n f)(x_j^{(n)}) = f(x_j^{(n)})$, $j = 0, \dots, n$, converges to f uniformly on $[a, b]$.*

Proof. The proof relies on the Weierstrass approximation theorem and the Chebyshev alternation theorem. The Weierstrass approximation theorem

(see [16]) ensures that for each $f \in C[a,b]$ there exists a sequence of polynomials $p_n \in P_n$ such that $\|p_n - f\|_\infty \to 0$ as $n \to \infty$. As a consequence of the Chebyshev alternation theorem from approximation theory (see [16]), for the uniquely determined best approximation $\tilde{p}_n$ to f in the maximum norm with respect to P_n, the error $\tilde{p}_n - f$ has at least $n+1$ zeros in $[a,b]$. Then taking the sequence of these zeros as the sequence of interpolation points implies the statement of the theorem. □

Theorem 8.17 (Faber) *For each sequence of interpolation points $(x_j^{(n)})$ there exists a function $f \in C[a,b]$ such that the sequence $(L_n f)$ of interpolation polynomials $L_n f \in P_n$ does not converge to f uniformly on $[a,b]$.*

Proof. This is a consequence of the uniform boundedness principle, Theorem 12.7. It implies that from the convergence of the sequence $(L_n f)$ for all $f \in C[a,b]$ it follows that there must exist a constant $C > 0$ such that $\|L_n\|_\infty \le C$ for all $n \in \mathbb{N}$. Then the statement of the theorem is obtained by showing that the interpolation operator L_n satisfies $\|L_n\|_\infty \ge c \ln n$ for all $n \in \mathbb{N}$ and some $c > 0$ (see [16]). □

We conclude this section by briefly describing Hermite interpolation, where in addition to the values of the polynomial, the values of its first derivative at the interpolation points are also prescribed.

Theorem 8.18 *Given $n+1$ distinct points $x_0, \ldots, x_n \in [a,b]$ and $2n+2$ values $y_0, \ldots, y_n \in \mathbb{R}$ and $y_0', \ldots, y_n' \in \mathbb{R}$, there exists a unique polynomial $p_{2n+1} \in P_{2n+1}$ with the property*

$$p_{2n+1}(x_j) = y_j, \quad p_{2n+1}'(x_j) = y_j', \quad j = 0, \ldots, n. \tag{8.10}$$

This Hermite interpolation polynomial *is given by*

$$p_{2n+1} = \sum_{k=0}^{n} [y_k H_k^0 + y_k' H_k^1] \tag{8.11}$$

with the Hermite factors

$$H_k^0(x) := [1 - 2\ell_k'(x_k)(x - x_k)]\,[\ell_k(x)]^2, \quad H_k^1(x) := (x - x_k)\,[\ell_k(x)]^2$$

expressed in terms of the Lagrange factors from Theorem 8.3.

Proof. Obviously, the polynomial p_{2n+1} belongs to P_{2n+1}, since the Hermite factors have degree $2n+1$. From (8.3), by elementary calculations it can be seen that (see Problem 8.7)

$$\begin{aligned} H_k^0(x_j) &= H_k^{1\prime}(x_j) = \delta_{jk}, \\ H_k^{0\prime}(x_j) &= H_k^1(x_j) = 0, \end{aligned} \qquad j,k = 0, \ldots, n. \tag{8.12}$$

From this it follows that the polynomial (8.11) satisfies the Hermite interpolation property (8.10).

To prove uniqueness of the Hermite interpolation polynomial we assume that $p_{2n+1,1}, p_{2n+1,2} \in P_{2n+1}$ are two polynomials having the interpolation property (8.10). Then the difference $p_{2n+1} := p_{2n+1,1} - p_{2n+1,2}$ satisfies

$$p_{2n+1}(x_j) = p'_{2n+1}(x_j) = 0, \quad j = 0, \dots, n;$$

i.e., the polynomial $p_{2n+1} \in P_{2n+1}$ has $n+1$ zeros of order two and therefore, by Theorem 8.1, must be identically equal to zero. This implies that $p_{2n+1,1} = p_{2n+1,2}$. □

The main application of Hermite interpolation consists in the approximation of a given function $f \in C^1[a,b]$ by interpolating its function values and the values of its derivative at $n+1$ distinct points $x_0, \dots, x_n \in [a,b]$. By

$$H_n : C^1[a,b] \to P_{2n+1}$$

we denote the *Hermite interpolation operator* that maps continuously differentiable functions $f : [a,b] \to \mathbb{R}$ into the uniquely determined Hermite interpolation polynomial $H_n f \in P_{2n+1}$ with the property

$$(H_n f)(x_j) = f(x_j), \quad (H_n f)'(x_j) = f'(x_j), \quad j = 0, \dots, n.$$

The following theorem can be proven analogously to Theorem 8.10 (see Problem 8.8).

Theorem 8.19 *Let $f : [a,b] \to \mathbb{R}$ be $(2n+2)$-times continuously differentiable. Then the remainder $R_n f := f - H_n f$ for Hermite interpolation with $n+1$ distinct points $x_0, \dots, x_n \in [a,b]$ can be represented in the form*

$$(R_n f)(x) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \prod_{j=0}^{n} (x - x_j)^2, \quad x \in [a,b], \tag{8.13}$$

for some $\xi \in [a,b]$ depending on x.

8.2 Trigonometric Interpolation

In applications, quite frequently there occur periodic functions, i.e., functions with the property

$$f(t+T) = f(t), \quad t \in \mathbb{R},$$

for some $T > 0$. For example, functions defined on closed planar or spatial curves always may be viewed as periodic functions. Polynomial interpolation is not appropriate for periodic functions, since algebraic polynomials

are not periodic. Therefore, we proceed by considering interpolation by trigonometric polynomials, which was first used independently by Clairaut (1759) and Lagrange (1762). Without loss of generality we assume that the period is equal to $T = 2\pi$.

Definition 8.20 *For $n \in \mathbb{N}$ we denote by T_n the linear space of* trigonometric polynomials

$$q(t) = \sum_{k=0}^{n} a_k \cos kt + \sum_{k=1}^{n} b_k \sin kt$$

with real (or complex) coefficients $a_0, \ldots, a_n$ and $b_1, \ldots, b_n$. A trigonometric polynomial $q \in T_n$ is said to be of degree n if $|a_n| + |b_n| > 0$.

From the addition theorems for the cosine and sine functions it follows that $q_1 q_2 \in T_{n_1+n_2}$ if $q_1 \in T_{n_1}$ and $q_2 \in T_{n_2}$. This justifies speaking of trigonometric polynomials.

Theorem 8.21 *A trigonometric polynomial in T_n that has more than $2n$ distinct zeros in the periodicity interval $[0, 2\pi)$ must vanish identically; i.e., all its coefficients must be equal to zero.*

Proof. We consider a trigonometric polynomial $q \in T_n$ of the form

$$q(t) = \frac{a_0}{2} + \sum_{k=1}^{n} [a_k \cos kt + b_k \sin kt]. \tag{8.14}$$

Setting $b_0 = 0$,

$$\gamma_k := \frac{1}{2}(a_k - ib_k), \quad \gamma_{-k} := \frac{1}{2}(a_k + ib_k), \quad k = 0, \ldots, n, \tag{8.15}$$

and using Euler's formula

$$e^{it} = \cos t + i \sin t,$$

we can rewrite (8.14) in the complex form

$$q(t) = \sum_{k=-n}^{n} \gamma_k e^{ikt}. \tag{8.16}$$

Therefore, substituting $z = e^{it}$ and setting

$$p(z) := \sum_{k=-n}^{n} \gamma_k z^{n+k},$$

we have the relation

$$q(t) = z^{-n} p(z).$$

Now assume that the trigonometric polynomial $q \in T_n$ has more than $2n$ distinct zeros in the interval $[0, 2\pi)$. Then the algebraic polynomial $p \in P_{2n}$ has more than $2n$ distinct zeros lying on the unit circle in the complex plane, since the function $t \mapsto e^{it}$ maps $[0, 2\pi)$ bijectively onto the unit circle. By Theorem 8.1, the algebraic polynomial p must be identically zero, and now (8.15) implies that also q must be identically zero. □

Theorem 8.22 *The cosine functions* $c_k(t) := \cos kt$, $k = 0, 1, \dots, n$, *and the sine functions* $s_k(t) := \sin kt$, $k = 1, \dots, n$, *are linearly independent in the function space* $C[0, 2\pi]$.

Proof. To prove this, assume that

$$\sum_{k=0}^{n} a_k c_k + \sum_{k=1}^{n} b_k s_k = 0;$$

that is,

$$\sum_{k=0}^{n} a_k \cos kt + \sum_{k=1}^{n} b_k \sin kt = 0, \quad t \in [0, 2\pi].$$

Then the trigonometric polynomial with coefficients $a_0, \dots, a_n$ and $b_1, \dots, b_n$ has more than $2n$ distinct zeros in $[0, 2\pi)$, and from Theorem 8.21 it follows that all the coefficients must be zero. Note that this linear independence also can be deduced from Theorem 3.17. □

Theorem 8.22 implies that the cosines c_k, $k = 0, 1, \dots, n$, and sines s_k, $k = 1, \dots, n$, form a basis for T_n and that T_n has dimension $2n + 1$.

Theorem 8.23 *Given* $2n+1$ *distinct points* $t_0, \dots, t_{2n} \in [0, 2\pi)$ *and* $2n+1$ *values* $y_0, \dots, y_{2n} \in \mathbb{R}$, *there exists a uniquely determined trigonometric polynomial* $q_n \in T_n$ *with the property*

$$q_n(t_j) = y_j, \quad j = 0, \dots, 2n. \tag{8.17}$$

In the Lagrange representation, *this* trigonometric interpolation polynomial *is given by*

$$q_n = \sum_{k=0}^{2n} y_k \ell_k \tag{8.18}$$

with the Lagrange factors

$$\ell_k(t) = \prod_{\substack{i=0 \\ i \neq k}}^{2n} \frac{\sin \dfrac{t - t_i}{2}}{\sin \dfrac{t_k - t_i}{2}}, \quad k = 0, \dots, 2n.$$

Proof. The function q_n belongs to T_n, since the Lagrange factors are trigonometric polynomials of degree n. The latter is a consequence of

$$\sin\frac{t-t_0}{2}\sin\frac{t-t_1}{2} = \frac{1}{2}\cos\frac{t_1-t_0}{2} - \frac{1}{2}\cos\left(t - \frac{t_1+t_0}{2}\right);$$

i.e., each of the functions ℓ_k is a product of n trigonometric polynomials of degree one. As in Theorem 8.3, we have $\ell_k(x_j) = \delta_{jk}$ for $j,k = 0,\dots,2n$, which shows that q_n indeed solves the trigonometric interpolation problem.

Uniqueness of the trigonometric interpolation polynomial follows analogously to the proof of Theorem 8.3 with the aid of Theorem 8.21. □

We now consider the important case of an equidistant subdivision

$$t_j = \frac{2\pi j}{2n+1}, \quad j = 0,\dots,2n.$$

For this we first note the summation formula

$$\sum_{j=0}^{2n} e^{ikt_j} = \sum_{j=0}^{2n} e^{ijt_k} = \begin{cases} 2n+1, & k = 0, \\ 0, & k = \pm 1,\dots,\pm 2n, \end{cases} \tag{8.19}$$

which is a consequence of the fact that for $e^{it_k} \neq 1$ we have the geometric sum

$$\sum_{j=0}^{2n} e^{ijt_k} = \frac{1 - e^{i(2n+1)t_k}}{1 - e^{it_k}} = 0,$$

whereas for $e^{it_k} = 1$ each term in the sum is equal to one.

We now attempt to find the uniquely determined interpolation polynomial in the complex form

$$q_n(t) = \sum_{k=-n}^{n} \gamma_k e^{ikt}.$$

From the interpolation conditions

$$q_n(t_j) = y_j, \quad j = 0,\dots,2n,$$

we observe that solving the interpolation problem is equivalent to solving the system of linear equations

$$\sum_{k=-n}^{n} \gamma_k \, e^{ikt_j} = y_j, \quad j = 0,\dots,2n. \tag{8.20}$$

Assume that the coefficients γ_k solve (8.20). Then, with the aid of (8.19), we obtain

$$\sum_{j=0}^{2n} y_j \, e^{-imt_j} = \sum_{k=-n}^{n} \gamma_k \sum_{j=0}^{2n} e^{i(k-m)t_j} = (2n+1)\gamma_m;$$

i.e., any solution of (8.20) must be of the form

$$\gamma_k = \frac{1}{2n+1} \sum_{j=0}^{2n} y_j \, e^{-ikt_j}, \quad k = -n, \ldots, n. \tag{8.21}$$

On the other hand, again with the aid of (8.19), for γ_k given by (8.21) we have that

$$\sum_{k=-n}^{n} \gamma_k \, e^{ikt_j} = \frac{1}{2n+1} \sum_{m=0}^{2n} y_m \sum_{k=0}^{2n} e^{ik(t_j - t_m)} = y_j, \quad j = 0, \ldots, 2n;$$

i.e., the linear system (8.20) has a unique solution, which is given by (8.21). From this, using the relation (8.15) between the real representation (8.14) and the complex representation (8.16) of trigonometric polynomials, we derive the following theorem.

Theorem 8.24 *There exists a unique trigonometric polynomial*

$$q_n(t) = \frac{a_0}{2} + \sum_{k=1}^{n} [a_k \cos kt + b_k \sin kt]$$

satisfying the interpolation property

$$q_n\left(\frac{2\pi j}{2n+1}\right) = y_j, \quad j = 0, \ldots, 2n.$$

Its coefficients are given by

$$a_k = \frac{2}{2n+1} \sum_{j=0}^{2n} y_j \cos \frac{2\pi jk}{2n+1}, \quad k = 0, \ldots, n,$$

$$b_k = \frac{2}{2n+1} \sum_{j=0}^{2n} y_j \sin \frac{2\pi jk}{2n+1}, \quad k = 1, \ldots, n.$$

For an equidistant subdivision with an even number $2n$ of interpolation points

$$t_j = \frac{\pi j}{n}, \quad j = 0, \ldots, 2n-1,$$

we have only $2n$ conditions to determine an element of the $(2n+1)$-dimensional space T_n. However, since the function $\sin nt$ obviously has its zeros at the interpolation points, we drop it from the interpolation polynomial. The proof of the following theorem is completely analogous to the proof of Theorem 8.24.

Theorem 8.25 *There exists a unique trigonometric polynomial*

$$q_n(t) = \frac{a_0}{2} + \sum_{k=1}^{n-1} [a_k \cos kt + b_k \sin kt] + \frac{a_n}{2} \cos nt$$

satisfying the interpolation property

$$q_n\left(\frac{\pi j}{n}\right) = y_j, \quad j = 0, \ldots, 2n-1.$$

Its coefficients are given by

$$a_k = \frac{1}{n} \sum_{j=0}^{2n-1} y_j \cos \frac{\pi j k}{n}, \quad k = 0, \ldots, n,$$

$$b_k = \frac{1}{n} \sum_{j=0}^{2n-1} y_j \sin \frac{\pi j k}{n}, \quad k = 1, \ldots, n-1.$$

Obviously, the trigonometric interpolation polynomials of Theorems 8.24 and 8.25 may be viewed as discretized versions of the Fourier series, where the integrals giving the coefficients of the Fourier series (see Problem 3.20) are approximated by the rectangular quadrature rule at an equidistant grid (see Corollary 9.27). Therefore, trigonometric interpolation on an equidistant grid is also known as the discrete Fourier transform.

An effective numerical evaluation of trigonometric polynomials can be done analogously to the Horner scheme for algebraic polynomials. For the polynomial

$$p(z) = \sum_{k=0}^{n} c_k z^k$$

the recursion (6.10) of the Horner scheme has the form

$$b_{k-1} = b_k z + c_{k-1}, \quad k = n, \ldots, 1,$$

starting with $b_n = c_n$, and it delivers $p(z) = b_0$. Assuming that the coefficients c_k are real, we substitute $z = e^{it}$ and separate into real and imaginary parts, $b_k = u_k + iv_k$, to obtain $u_n = c_n$, $v_n = 0$, and the recursion

$$u_{k-1} = u_k \cos t - v_k \sin t + c_{k-1}, \quad v_{k-1} = u_k \sin t + v_k \cos t,$$

for $k = n-1, \ldots, 1$. From this we find

$$u_0 = \sum_{k=0}^{n} c_k \cos kt, \quad v_0 = \sum_{k=1}^{n} c_k \sin kt;$$

i.e., the evaluation of a trigonometric polynomial at a point t can be reduced to the evaluation of $\sin t$ and $\cos t$ and $O(n)$ additions and multiplications.

To compute all the coefficients a_k and b_k in Theorem 8.24 or 8.25 by this approach requires $O(n^2)$ additions and multiplications.

By the *fast Fourier transform*, which is attributed to Cooley and Tukey (1965) and which was known already to Gauss, the computational costs can be reduced even further. The main idea is to exploit the symmetries of $e^{2\pi i j/n}$ if n is a power of two, say $n = 2^p$ with $p \in \mathbb{N}$. We briefly explain the fast Fourier transform algorithm for the evaluation of the *discrete Fourier transform* in the complex form

$$c_k = \frac{1}{n} \sum_{j=0}^{n-1} y_j e^{-\frac{2\pi i}{n} kj}, \quad k = 0, \ldots, n-1. \tag{8.22}$$

Let $m := n/2 = 2^{p-1}$ and $\omega := e^{-2\pi i/n}$. Then $\omega^n = 1$, $\omega^m = -1$, and (8.22) reads

$$c_k = \frac{1}{n} \sum_{j=0}^{n-1} y_j \omega^{kj}, \quad k = 0, \ldots, n-1.$$

Now, the basic idea of the Cooley–Tukey algorithm is to break this sum into two parts for j even and j odd; i.e.,

$$c_k = \frac{1}{2}\, \gamma_k + \frac{1}{2}\, \omega^k \delta_k, \quad k = 0, \ldots, n-1,$$

where

$$\gamma_k := \frac{1}{m} \sum_{j=0}^{m-1} y_{2j} \omega^{2jk}, \quad \delta_k := \frac{1}{m} \sum_{j=0}^{m-1} y_{2j+1} \omega^{2jk}, \quad k = 0, \ldots, n-1.$$

Since $\omega^2 = e^{-2\pi i/m}$, we have $\gamma_{k+m} = \gamma_k$ and $\delta_{k+m} = \delta_k$, and therefore

$$c_k = \frac{1}{2}\, \gamma_k + \frac{1}{2}\, \omega^k \delta_k, \quad c_{k+m} = \frac{1}{2}\, \gamma_k - \frac{1}{2}\, \omega^k \delta_k, \quad k = 0, \ldots, m-1.$$

Obviously, the γ_k, δ_k, $k = 0, \ldots, m-1$, represent a discrete Fourier transform of length $m = n/2$. Hence, the discrete Fourier transform of length n is reduced to two discrete Fourier transforms of length $n/2$ followed by n multiplications and n additions. If this is done recursively, we arrive at the following operation count. Assuming that the ω^k, $k = 1, \ldots, m-1$, are precomputed, let M_p denote the number of additions and multiplications needed for the Fourier transform of length $n = 2^p$. Then,

$$M_p = 2M_{p-1} + 2^{p+1}$$

with $M_0 = 0$. From this, by induction, it follows that

$$M_p = p\, 2^{p+1} = 2n \log_2 n,$$

i.e., that the computational cost is reduced significantly from order $O(n^2)$ to order $O(n \log_2 n)$.

The actual numerical implementation is based on writing the indices k and j in a binary representation

$$k = [k_0, \dots, k_{p-1}] = \sum_{q=0}^{p-1} k_q 2^q, \quad j = [j_0, \dots, j_{p-1}] = \sum_{q=0}^{p-1} j_q 2^q$$

with $k_q, j_q \in \{0, 1\}$ for $q = 0, \dots, p-1$. Then

$$e^{-\frac{2\pi i}{n} kj} = \prod_{q=0}^{p-1} e^{-\frac{2\pi i j_q}{2^{p-q}} [k_0, \dots, k_{p-q-1}]},$$

since

$$e^{-2\pi i j_q k_r 2^{q+r-p}} = 1$$

for $q + r \geq p$. Inserting this into (8.22), we can split the long sum into p nested short sums, and the Fourier transform becomes

$$c_k = \frac{1}{n} \sum_{j_0=0}^{1} e^{-\frac{2\pi i j_0}{2^p} [k_0, \dots, k_{p-1}]} \times \sum_{j_1=0}^{1} e^{-\frac{2\pi i j_1}{2^{p-1}} [k_0, \dots, k_{p-2}]}$$

$$\times \cdots \times \sum_{j_{p-1}=0}^{1} e^{-\frac{2\pi i j_{p-1}}{2} k_0} \, y_{[j_0, \dots, j_{p-1}]}.$$

Define the intermediate sums

$$S^q_{[j_0, \dots, j_{p-q-1}, k_{q-1}, \dots, k_0]} := \sum_{j_{p-q}=0}^{1} e^{-\frac{2\pi i j_{p-q}}{2^{p-q}} [k_0, \dots, k_{q-1}]}$$

$$\times \cdots \times \sum_{j_{p-1}=0}^{1} e^{-\frac{2\pi i j_{p-1}}{2} k_0} \, y_{[j_0, \dots, j_{p-1}]}$$

for $q = 1, \dots, p$ and $j_0, \dots, j_{p-q-1}, k_{q-1}, \dots, k_0 \in \{0, 1\}$. Then clearly,

$$c_{[k_0, \dots, k_{p-1}]} = \frac{1}{n} S^p_{[k_{p-1}, \dots, k_0]}, \tag{8.23}$$

and setting

$$S^0_{[j_0, \dots, j_{p-1}]} = y_{[j_0, \dots, j_{p-1}]},$$

we have the recursive relation

$$S^q_{[j_0, \dots, j_{p-q-1}, k_{q-1}, \dots, k_0]} = S^{q-1}_{[j_0, \dots, j_{p-q-1}, 0, k_{q-2}, \dots, k_0]}$$

$$+ S^{q-1}_{[j_0, \dots, j_{p-q-1}, 1, k_{q-2}, \dots, k_0]} e^{-\frac{2\pi i}{2^{p-q}} [k_0, \dots, k_{q-1}]}$$

for $q = 1, \dots, p$. Each step of these p recursions requires n additions and n multiplications. Hence, the total computational cost is indeed of order $O(n \log_2 n)$. For more details of the actual numerical implementation and, in particular, on how to effectively perform the so-called bit reversal in order to arrange the result (8.23) in the natural order, we refer to [45].

The error analysis for trigonometric interpolation is more complicated than the error analysis of the previous section for polynomial interpolation. Denote by $L_n : C[0, 2\pi] \to T_n$ the *trigonometric interpolation operator* that maps the function f onto its trigonometric interpolation polynomial $L_n f$. For equidistant grids, by Problems 8.12 and 8.13 we have convergence $\|L_n f - f\|_2 \to 0$, $n \to \infty$, for each continuous 2π-periodic function f and $\|L_n f - f\|_\infty \to 0$, $n \to \infty$, for each continuously differentiable 2π-periodic function f. For a detailed error analysis we refer to [49].

8.3 Spline Interpolation

As we have seen in our considerations of the convergence of interpolation polynomials, increasing the number of interpolation points, i.e., increasing the degree of the polynomials, does not always lead to an improvement in the approximation. The *spline interpolation* that we will study in this section remedies this deficiency of interpolation by high-degree polynomials through a piecewise polynomial interpolation of low degree.

A frequently used method of this type is piecewise linear interpolation. Let $a = x_0 < x_1 < \cdots < x_n = b$ be a subdivision of the interval $[a, b]$. Then a given function $f \in C[a, b]$ can be approximated by a continuous piecewise linear function by linear interpolation on each of the subintervals, i.e., according to Example 8.12, by

$$s_n(x) = \frac{1}{x_j - x_{j-1}} \left[f(x_{j-1})(x_j - x) + f(x_j)(x - x_{j-1}) \right], \quad x \in [x_{j-1}, x_j].$$

From the error estimate (8.9) for linear interpolation, we see that for piecewise linear interpolation we have uniform convergence $\|s_n - f\|_\infty \to 0$ for $n \to \infty$ on $[a, b]$, provided that $h := \max_{j=1,\dots,n} |x_j - x_{j-1}| \to 0$ and $f \in C^2[a, b]$. The main advantage of this method is its simplicity and its stability with respect to errors in the interpolation values. However, since by (8.9) linear interpolation has an error only of order $O(h^2)$, for achieving a prescribed accuracy it usually requires a much finer discretization than some of the higher-order methods described below.

Definition 8.26 *Let $a = x_0 < x_1 < \cdots < x_n = b$ be a subdivision of the interval $[a, b]$ and $m \in \mathbb{N}$. A function $s : [a, b] \to \mathbb{R}$ is called a* spline of degree *m with respect to this subdivision if s is $(m-1)$-times continuously differentiable on $[a, b]$ and if the restriction of s to each subinterval*

$[x_{j-1}, x_j]$ for $j = 1, \dots, n$ reduces to a polynomial of degree at most m. By S_m^n we denote the set of all splines of degree m for a fixed subdivision.

Although piecewise polynomials have been studied since the beginning of this century, the notation *spline* was introduced only in 1946 by Schoenberg. The term originates from the thin wooden or metal strips that were used by draftsmen to fit a smooth curve between specified points. Since small displacements s of a thin elastic beam are governed by the fourth-order differential equation $s^{(4)} = 0$, *cubic splines*, i.e., splines of degree three, indeed model the draftsmen's splines.

Theorem 8.27 *S_m^n is a linear space of dimension $m + n$.*

Proof. Clearly, S_m^n is a linear space, since $C^{m-1}[a, b]$ and P_m are linear spaces. In the sequel we shall use the notation

$$x_+^m := \begin{cases} x^m, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

for $m \in \mathbb{N}$. The $m + n$ functions

$$\begin{aligned} u_k(x) &:= (x - x_0)^k, \quad k = 0, \dots, m, \\ v_k(x) &:= (x - x_k)_+^m, \quad k = 1, \dots, n - 1, \end{aligned} \tag{8.24}$$

are linearly independent. In order to see this, let

$$\sum_{k=0}^{m} \alpha_k u_k + \sum_{k=1}^{n-1} \beta_k v_k = 0.$$

Then, in particular,

$$\sum_{k=0}^{m} \alpha_k (x - x_0)^k = 0, \quad x \in [x_0, x_1],$$

whence $\alpha_k = 0$ for $k = 0, \dots, m$. Then we have

$$\beta_1 (x - x_1)^m = 0, \quad x \in [x_1, x_2],$$

and therefore $\beta_1 = 0$. Repeating this argument inductively, it follows that $\beta_k = 0$, $k = 1, \dots, n - 1$.

To complete the proof, we need to show that each $s \in S_m^n$ can be expressed as a linear combination of the functions (8.24). Given a spline $s \in S_m^n$, by induction we show that there exist constants $\alpha_0, \dots, \alpha_m$ and $\beta_1, \dots, \beta_{n-1}$ such that

$$s(x) = \sum_{k=0}^{m} \alpha_k (x - x_0)^k + \sum_{k=1}^{j-1} \beta_k (x - x_k)_+^m, \quad x \in [x_0, x_j], \tag{8.25}$$

for $j = 1, \ldots, n$. This is true for $j = 1$, since on $[x_0, x_1]$ the spline s coincides with an element of P_m. Now assume that we have the representation (8.25) for some $j \geq 1$. Then the difference

$$p(x) := s(x) - \sum_{k=0}^{m} \alpha_k (x - x_0)^k - \sum_{k=1}^{j-1} \beta_k (x - x_k)_+^m$$

restricted to the interval $[x_j, x_{j+1}]$ is in P_m. Since the spline s is in $C^{m-1}[a, b]$ and p vanishes on $[x_0, x_j]$ we have that

$$p^{(i)}(x_j) = 0, \quad i = 0, \ldots, m - 1.$$

Hence $p(x) = \beta_j (x - x_j)_+^m$ on $[x_j, x_{j+1}]$ for some constant β_j, and because $(x - x_j)_+^m = 0$ on $[x_0, x_j]$, the representation (8.25) is proven for $j + 1$. □

Since the spline space S_m^n has dimension $m + n$, the $n + 1$ interpolation conditions at the points $x_0, \ldots, x_n$ are not sufficient to determine uniquely a spline of degree greater than one. Therefore, we need to add additional requirements in the form of conditions at the two endpoints $x_0 = a$ and $x_n = b$ of the interval. Since we want to divide the number of these end conditions equally between both ends, we consider only odd degrees m.

Lemma 8.28 *Let $m = 2\ell - 1$ with $\ell \in \mathbb{N}$ and $\ell \geq 2$, and let $f \in C^\ell[a, b]$. Assume that the spline $s \in S_m^n$ interpolates f, i.e.,*

$$s(x_j) = f(x_j), \quad j = 0, \ldots, n, \tag{8.26}$$

and that it satisfies the boundary conditions

$$s^{(j)}(a) = f^{(j)}(a), \quad s^{(j)}(b) = f^{(j)}(b), \quad j = 1, \ldots, \ell - 1. \tag{8.27}$$

Then

$$\int_a^b [f^{(\ell)}(x) - s^{(\ell)}(x)]^2 dx = \int_a^b [f^{(\ell)}(x)]^2 dx - \int_a^b [s^{(\ell)}(x)]^2 dx. \tag{8.28}$$

Proof. We have that

$$\int_a^b [f^{(\ell)}(x) - s^{(\ell)}(x)]^2 dx = \int_a^b [f^{(\ell)}(x)]^2 dx - \int_a^b [s^{(\ell)}(x)]^2 dx - 2R,$$

where

$$R := \int_a^b [f^{(\ell)}(x) - s^{(\ell)}(x)] s^{(\ell)}(x)\, dx.$$

Since $f \in C^\ell[a, b]$ and $s \in C^{m-1}[a, b]$ has piecewise continuous derivatives of order m, by $\ell - 1$ repeated partial integrations and using the boundary conditions (8.27) we obtain that

$$R = (-1)^{\ell-1} \int_a^b [f'(x) - s'(x)] s^{(m)}(x)\, dx.$$

A further partial integration and the interpolation conditions now yield

$$R = (-1)^{\ell-1} \sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} [f'(x) - s'(x)] s^{(m)}(x)\, dx$$

$$= (-1)^{\ell-1} \sum_{j=1}^{n} [f(x) - s(x)] s^{(m)}(x) \Big|_{x_{j-1}}^{x_j} = 0,$$

since $s^{(m+1)} = 0$. This completes the proof. □

Lemma 8.29 *Under the assumptions of Lemma 8.28 let $f = 0$. Then $s = 0$.*

Proof. For $f = 0$, from (8.28) it follows that

$$\int_a^b [s^{(\ell)}(x)]^2 dx = 0.$$

This implies that $s^{(\ell)} = 0$, and therefore $s \in P_{\ell-1}$ on $[a, b]$. Now the boundary conditions $s^{(j)}(a) = 0$, $j = 0, \ldots, \ell - 1$, yield $s = 0$. □

From the proof it can be seen that Lemmas 8.28 and 8.29 remain valid if the boundary conditions (8.27) are replaced by

$$s^{(\ell+j)}(a) = s^{(\ell+j)}(b) = 0, \quad j = 0, \ldots, \ell - 2,$$

or, provided that f is periodic with period $b-a$, by the periodicity condition

$$s^{(j)}(a) = s^{(j)}(b), \quad j = 1, \ldots, \ell - 1.$$

Consequently, the following conclusions drawn from Lemma 8.29 are also true for these two end conditions. However, from a practical point of view only the latter modification is of relevance.

Theorem 8.30 *Let $m = 2\ell - 1$ with $\ell \in \mathbb{N}$ and $\ell \geq 2$. Then, given $n + 1$ values $y_0, \ldots, y_n$ and $m - 1$ boundary data $a_1, \ldots, a_{\ell-1}$ and $b_1, \ldots, b_{\ell-1}$, there exists a unique spline $s \in S_m^n$ satisfying the interpolation conditions*

$$s(x_j) = y_j, \quad j = 0, \ldots, n, \tag{8.29}$$

and the boundary conditions

$$s^{(j)}(a) = a_j, \quad s^{(j)}(b) = b_j, \quad j = 1, \ldots, \ell - 1. \tag{8.30}$$

Proof. Representing the spline in the form (8.25), i.e.,

$$s(x) = \sum_{k=0}^{m} \alpha_k u_k + \sum_{k=1}^{n-1} \beta_k v_k, \tag{8.31}$$

it follows that the interpolation conditions (8.29) and boundary conditions (8.30) are satisfied if and only if the $m+n$ coefficients $\alpha_0,\dots,\alpha_m$ and $\beta_1,\dots,\beta_{n-1}$ solve the system

$$\begin{aligned}
\sum_{k=0}^{m}\alpha_k u_k(x_j)+\sum_{k=1}^{n-1}\beta_k v_k(x_j) &= y_j, \quad j=0,\dots,n,\\
\sum_{k=0}^{m}\alpha_k u_k^{(j)}(a)+\sum_{k=1}^{n-1}\beta_k v_k^{(j)}(a) &= a_j, \quad j=1,\dots,\ell-1,\\
\sum_{k=0}^{m}\alpha_k u_k^{(j)}(b)+\sum_{k=1}^{n-1}\beta_k v_k^{(j)}(b) &= b_j, \quad j=1,\dots,\ell-1,
\end{aligned} \tag{8.32}$$

of $m+n$ linear equations. By Lemma 8.29 the homogeneous form of the system (8.32) has only the trivial solution. Therefore, the inhomogeneous system (8.32) is uniquely solvable, and the proof is finished. □

In principle, for the actual computation of the interpolating spline, it is possible to use the linear system (8.32). However, as a consequence of the global nature of the basis functions (8.24), this system turns out to be ill-conditioned. Therefore, it is preferable to use the corresponding linear system derived from another set of basis functions known as basic splines, or simply *B-splines*. As opposed to the splines (8.24), the B-splines have local support, i.e., they differ from zero only within $m+1$ neighboring subintervals.

For the sake of simplicity we confine our analysis of B-splines to the case of an equidistant subdivision of step length h. We set

$$B_0(x) := \begin{cases} 1, & |x| \le 0.5,\\ 0, & |x| > 0.5,\end{cases}$$

and define recursively

$$B_{m+1}(x) := \int_{x-\frac{1}{2}}^{x+\frac{1}{2}} B_m(y)\,dy, \quad x\in\mathbb{R}, \quad m=0,1,\dots. \tag{8.33}$$

Then, by induction, it can be seen that the B_m are $(m-1)$-times continuously differentiable and nonnegative, vanish outside the interval $[-m/2-1/2, m/2+1/2]$, and reduce to a polynomial of degree m in each of the intervals $[i,i+1]$ for m odd and $[i-1/2,i+1/2]$ for m even for i an integer; i.e., the B_m are splines of order m on an integer grid if m is odd and on a half integer grid if m is even.

Elementary integrations show that

$$B_1(x) = \begin{cases} 1 - |x|, & |x| \le 1, \\ 0, & |x| \ge 1, \end{cases} \tag{8.34}$$

$$B_2(x) = \frac{1}{2}\begin{cases} 2 - (|x| - 0.5)^2 - (|x| + 0.5))^2, & |x| \le 0.5, \\ (|x| - 1.5)^2, & 0.5 \le |x| \le 1.5, \\ 0, & |x| \ge 1.5, \end{cases} \tag{8.35}$$

$$B_3(x) = \frac{1}{6}\begin{cases} (2 - |x|)^3 - 4(1 - |x|)^3, & |x| \le 1, \\ (2 - |x|)^3, & 1 \le |x| \le 2, \\ 0, & |x| \ge 2. \end{cases} \tag{8.36}$$

Graphs of these B-splines are given in Figure 8.1.

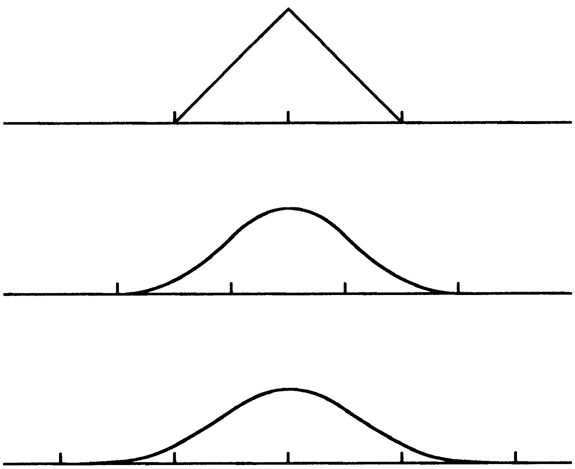


FIGURE 8.1. B-splines B_1, B_2, and B_3

Theorem 8.31 *For $m \in \mathbb{N} \cup \{0\}$ the B-splines*

$$B_m(\cdot - k), \quad k = 0, \dots, m, \tag{8.37}$$

are linearly independent on the interval $I_m := \left[\frac{m-1}{2}, \frac{m+1}{2}\right]$.

Proof. This is trivial for $m = 0$, and we assume that it has been proven for degree $m-1$ for some $m \geq 1$. Let

$$\sum_{k=0}^{m} \alpha_k B_m(x-k) = 0, \quad x \in I_m. \tag{8.38}$$

Then, with the aid of (8.33), differentiating (8.38) yields

$$\sum_{k=0}^{m} \alpha_k \left[B_{m-1}\left(x-k+\frac{1}{2}\right) - B_{m-1}\left(x-k-\frac{1}{2}\right)\right] = 0, \quad x \in I_m.$$

Observing that the supports of $B_{m-1}\left(\cdot + \frac{1}{2}\right)$ and $B_{m-1}\left(\cdot - m - \frac{1}{2}\right)$ do not intersect with I_m, we can rewrite this as

$$\sum_{k=1}^{m} [\alpha_k - \alpha_{k-1}] B_{m-1}\left(x-k+\frac{1}{2}\right) = 0, \quad x \in I_m,$$

whence $\alpha_k = \alpha_{k-1}$ for $k = 1, \ldots, m$ follows by the induction assumption; i.e., $\alpha_k = \alpha$ for $k = 0, \ldots, m$. Now (8.38) reads

$$\alpha \sum_{k=0}^{m} B_m(x-k) = 0, \quad x \in I_m,$$

and integrating this equation over the interval I_m leads to

$$\alpha \int_{-\frac{m}{2}-\frac{1}{2}}^{\frac{m}{2}+\frac{1}{2}} B_m(x)\, dx = 0.$$

This finally implies $\alpha = 0$, since the B_m are nonnegative, and the proof is finished. □

Corollary 8.32 *Let $x_k = a + hk$, $k = 0, \ldots, n$, be an equidistant subdivision of the interval $[a,b]$ of step size $h = (b-a)/n$ with $n \geq 2$, and let $m = 2\ell - 1$ with $\ell \in \mathbb{N}$. Then the B-splines*

$$B_{m,k}(x) := B_m\left(\frac{x-a-hk}{h}\right), \quad x \in [a,b], \tag{8.39}$$

for $k = -\ell+1, \ldots, n+\ell-1$ form a basis for S_m^n.

Proof. The $n+m$ splines (8.39) belong to S_m^n, and by the preceding Theorem 8.31 they can be shown to be linearly independent on $[a,b]$. Hence, the statement follows from Theorem 8.27. □

The use of the B-splines as a basis opens up another possibility for the computation of an interpolating spline. We only consider the case $m = 3$, i.e., *cubic splines.* From (8.36) we note that

$$B_3(0) = \frac{2}{3}, \quad B_3(\pm 1) = \frac{1}{6}, \quad B_3'(0) = 0, \quad B_3'(\pm 1) = \mp\frac{1}{2}.$$

Therefore, the cubic spline

$$s(x) = \sum_{k=-1}^{n+1} \alpha_k \, B_3\left(\frac{x - x_k}{h}\right), \quad x \in [a,b], \tag{8.40}$$

satisfies the interpolation conditions (8.29) and the boundary conditions (8.30) if and only if the $n+3$ coefficients $\alpha_{-1}, \dots, \alpha_{n+1}$ satisfy the system

$$\begin{aligned} -\frac{1}{2}\,\alpha_{-1} + \frac{1}{2}\,\alpha_1 &= ha_1, \\ \frac{1}{6}\,\alpha_{j-1} + \frac{2}{3}\,\alpha_j + \frac{1}{6}\,\alpha_{j+1} &= y_j, \quad j = 0, \dots, n, \\ -\frac{1}{2}\,\alpha_{n-1} + \frac{1}{2}\,\alpha_{n+1} &= hb_1, \end{aligned} \tag{8.41}$$

of $n+3$ linear equations. Since the matrix of this system is irreducible and weakly row-diagonally dominant, the solution can be obtained by Jacobi iteration (see Theorem 4.7).

We conclude this section with an analysis of the interpolation error for cubic splines and note that the results can be extended to arbitrary odd degree. We begin with a convergence result for arbitrary subdivisions under a weak regularity assumption on the interpolated function.

Theorem 8.33 *Let $f : [a,b] \to \mathbb{R}$ be twice continuously differentiable and let $s \in S_3^n$ be the uniquely determined cubic spline satisfying the interpolation and boundary conditions of Lemma 8.28. Then*

$$\|f - s\|_\infty \le \frac{h^{3/2}}{2}\,\|f''\|_2 \quad \text{and} \quad \|f' - s'\|_\infty \le h^{1/2}\,\|f''\|_2,$$

where $h := \max_{j=1,\dots,n} |x_j - x_{j-1}|$.

Proof. The error function $r := f - s$ has $n+1$ zeros $x_0, \dots, x_n$. Hence, the distance between two consecutive zeros of r is less than or equal to h. By Rolle's theorem, the derivative r' has n zeros with distance less than or equal to $2h$. Choose $z \in [a,b]$ such that $|r'(z)| = \|r'\|_\infty$. Then the closest zero ζ of r' has distance $|\zeta - z| \le h$, and by the Cauchy–Schwarz inequality we can estimate

$$\|r'\|_\infty^2 = \left|\int_\zeta^z r''(y)\,dy\right|^2 \le h\left|\int_\zeta^z [r''(y)]^2 dy\right| \le h\int_a^b [r''(y)]^2 dy.$$

From this, using Lemma 8.28 we obtain $\|r'\|_\infty \le \sqrt{h}\,\|f''\|_2$.

Choose $x \in [a,b]$ such that $|r(x)| = \|r\|_\infty$. Then the closest zero ξ of r has distance $|\xi - x| \le h/2$, and we can estimate

$$\|r\|_\infty = \left| \int_\xi^x r'(y)dy \right| \le \frac{h}{2}\, \|r'\|_\infty \le \frac{h\sqrt{h}}{2}\, \|f''\|_2,$$

which concludes the proof. □

If we assume more regularity on f, we can improve on the order of convergence. For this we need to derive an estimate on the second derivative of the interpolating spline. From (8.36) it follows that

$$B_3''(0) = -2 \quad \text{and} \quad B_3''(\pm 1) = 1.$$

Hence, the cubic spline (8.40) has second derivatives given by the difference formula

$$s''(x_j) = \frac{1}{h^2}\, [\alpha_{j-1} - 2\alpha_j + \alpha_{j+1}], \quad j = 0, \dots, n. \tag{8.42}$$

From this we deduce that

$$\begin{aligned} h^2[s''(x_{j-1}) + 4s''(x_j) + s''(x_{j+1})] = {} & [\alpha_{j-2} + 4\alpha_{j-1} + \alpha_j] \\ & -2[\alpha_{j-1} + 4\alpha_j + \alpha_{j+1}] \\ & +[\alpha_j + 4\alpha_{j+1} + \alpha_{j+2}] \end{aligned}$$

for $j = 1, \dots, n-1$,

$$\begin{aligned} h^2[4s''(x_0) + 2s''(x_1)] = {} & 6[\alpha_{-1} - \alpha_1] - 2[\alpha_{-1} + 4\alpha_0 + \alpha_1] \\ & +2[\alpha_0 + 4\alpha_1 + \alpha_2], \end{aligned}$$

and

$$\begin{aligned} h^2[2s''(x_{n-1}) + 4s''(x_n)] = {} & 2[\alpha_{n-2} + 4\alpha_{n-1} + \alpha_n] \\ & -2[\alpha_{n-1} + 4\alpha_n + \alpha_{n+1}] - 6[\alpha_{n-1} - \alpha_{n+1}]. \end{aligned}$$

From this and the linear system (8.41), for the special case of the interpolation conditions (8.26) and the boundary conditions (8.27), it follows that the $n+1$ values of s'' at the grid points satisfy the system

$$\begin{aligned} 4s''(x_0) + 2s''(x_1) &= F_0, \\ s''(x_{j-1}) + 4s''(x_j) + s''(x_{j+1}) &= F_j, \quad j = 1, \dots, n-1, \\ 2s''(x_{n-1}) + 4s''(x_n) &= F_n, \end{aligned} \tag{8.43}$$

of $n + 1$ linear equations with right-hand sides

$$F_0 := \frac{12}{h^2}\,[-f(x_0) + f(x_1) - hf'(x_0)],$$

$$F_j := \frac{6}{h^2}\,[f(x_{j-1}) - 2f(x_j) + f(x_{j+1})], \quad j = 1, \ldots, n-1,$$

$$F_n := \frac{12}{h^2}\,[f(x_{n-1}) - f(x_n) + hf'(x_n)].$$

From the system (8.43) we can conclude that

$$4|s''(x_j)| \le |F_j| - 2 \max_{k=0,\ldots,n} |s''(x_k)|, \quad j = 0, \ldots, n,$$

and therefore

$$\max_{j=0,\ldots,n} |s''(x_j)| \le \frac{1}{2} \max_{j=0,\ldots,n} |F_j|. \tag{8.44}$$

If f is twice continuously differentiable, by Taylor's formula we can estimate

$$\max\{|F_0|, |F_n|\} \le 6\,\|f''\|_\infty.$$

From Example 8.12, applied to the remainder in the linear interpolation of $f(x_j)$ from $f(x_{j-1})$ and $f(x_{j+1})$, we obtain

$$|F_j| = \frac{6}{h^2}|f(x_{j-1}) - 2f(x_j) + f(x_{j+1})| \le 6\|f''\|_\infty, \quad j = 1, \ldots, n-1.$$

Hence, since s'' is piecewise linear, from (8.44) it follows that

$$\|s''\|_\infty \le 3\|f''\|_\infty. \tag{8.45}$$

Theorem 8.34 *Let $f : [a, b] \to \mathbb{R}$ be four-times continuously differentiable and let $s \in S_3^n$ be the uniquely determined cubic spline satisfying the interpolation and boundary conditions of Lemma 8.28 for an equidistant subdivision with step width h. Then*

$$\|f - s\|_\infty \le \frac{h^4}{16}\,\|f^{(4)}\|_\infty.$$

Proof. By $L_1 : C[a, b] \to S_1^n$ we denote the interpolation operator mapping $g \in C[a, b]$ onto its uniquely determined piecewise linear interpolation. From Example 8.12 we obtain that

$$\|r\|_\infty = \|r - L_1 r\|_\infty \le \frac{h^2}{8}\,\|r''\|_\infty,$$

since trivially $L_1 r = 0$.

By integration, we choose a function w such that $w'' = L_1 f''$. Applying the estimate (8.45) for the cubic spline $s - w$ and using the estimate (8.9) for the piecewise linear interpolation of f'', we obtain

$$\|f''-s''\|_\infty \leq \|f''-L_1 f''\|_\infty + \|L_1 f''-s''\|_\infty \leq 4\|f''-L_1 f''\|_\infty \leq \frac{h^2}{2}\,\|f^{(4)}\|_\infty.$$

By piecing together the last two inequalities we obtain the assertion of the theorem. □

8.4 Bézier Polynomials

In this section we want to introduce some of the basic ideas of *computer-aided geometric design.* We will confine our presentation to planar (and spatial) curves, i.e., to subsets $\Gamma \subset \mathbb{R}^m$, $m = 2, 3$, that can be described by a continuous mapping $f : D \to \mathbb{R}^m$ of an interval $D \subset \mathbb{R}$ into $\mathbb{R}^m$. For the purposes of computer-aided geometric design it is essential that the geometric objects can be visualized and manipulated on the computer very effectively and rapidly. This, in particular, makes it essential that the parameters entering the representation of the curves have a geometric meaning. The latter property, for example, is not fulfilled by polynomial curves represented through the classical monomial basis.

Definition 8.35 *For $n \in \mathbb{N} \cup \{0\}$, we denote by P_n^m the linear space of polynomials of the form*

$$p(x) = \sum_{k=0}^{n} a_k x^k, \quad x \in \mathbb{R},$$

where $a_0, \ldots, a_n \in \mathbb{R}^m$. A polynomial $p \in P_n^m$ is said to be of degree n if $a_n \neq 0$.

We proceed by introducing a basis for polynomials on an interval $[a, b]$ in $\mathbb{R}$ with $a < b$ that is better suited for the purposes of computer-aided design than the monomial basis. For this we make use of the fact that by the affine linear transformation

$$x \mapsto t(x) := \frac{x-a}{b-a} \tag{8.46}$$

the interval $[a, b]$ can be mapped on the interval $[0, 1]$. By the binomial formula we have that

$$1 = [t + (1-t)]^n = \sum_{k=0}^{n} \binom{n}{k} t^k (1-t)^{n-k}.$$

The terms in this partition of unity are called Bernstein polynomials for the interval $[0, 1]$. From these, the Bernstein polynomials for the interval $[a, b]$ are obtained via the transformation (8.46).

Definition 8.36 *The* Bernstein polynomials *of degree n for the interval $[0,1]$ are given by*

$$B_k^n(t) := \binom{n}{k} t^k (1-t)^{n-k}, \quad k = 0, \dots, n. \tag{8.47}$$

Correspondingly, the polynomials

$$B_k^n(x; a, b) := B_k^n\left(\frac{x-a}{b-a}\right) = \frac{1}{(b-a)^n}\binom{n}{k}(x-a)^k(b-x)^{n-k}, \quad k = 0, \dots, n,$$

are called Bernstein polynomials of degree n for the interval $[a,b]$.

Some basic properties of Bernstein polynomials are described in the following theorem.

Theorem 8.37 *The Bernstein polynomials are nonnegative on $[0,1]$ and provide a* partition of unity; *i.e.,*

$$B_k^n(t) \geq 0, \quad t \in [0,1], \tag{8.48}$$

and

$$\sum_{k=0}^{n} B_k^n(t) = 1, \quad t \in \mathbb{R}. \tag{8.49}$$

They satisfy the relations

$$B_k^n(t) = B_{n-k}^n(1-t), \quad k = 0, \dots, n, \tag{8.50}$$

and

$$B_0^n(t) = (1-t)B_0^{n-1}(t), \quad B_n^n(t) = tB_{n-1}^{n-1}(t) \tag{8.51}$$

for all $t \in \mathbb{R}$ and $n \in \mathbb{N}$. The point $t = 0$ is a zero of B_k^n of order k, and $t = 1$ is a zero of order $n - k$. Each of the polynomials B_k^n assumes its maximum value only at $t = k/n$. They satisfy the recursion relation

$$B_k^n(t) = tB_{k-1}^{n-1}(t) + (1-t)B_k^{n-1}(t), \quad t \in \mathbb{R}, \tag{8.52}$$

for $n \in \mathbb{N}$ and $k = 1, \dots, n-1$. The polynomials $B_0^n, \dots, B_n^n$ form a basis of P_n.

Proof. The first five properties are obvious. The statement on the maximum of B_k^n is a consequence of

$$\frac{d}{dt} B_k^n(t) = \binom{n}{k} t^{k-1}(1-t)^{n-k-1}(k - nt), \quad k = 0, \dots n.$$

The recursion formula (8.52) follows from the definition (8.47) and the recursion formula

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

for the binomial coefficients. In order to show that the $n+1$ polynomials $B_0^n, \dots, B_n^n$ of degree n provide a basis of P_n, we prove that they are linearly independent. Let

$$\sum_{k=0}^{n} b_k B_k^n(t) = 0, \quad t \in [0,1].$$

Then

$$\sum_{k=0}^{n} b_k \frac{d^j}{dt^j} B_k^n(t) = 0, \quad t \in [0,1],$$

and therefore

$$\sum_{k=j}^{n} b_k \frac{d^j}{dt^j} B_k^n(0) = 0, \quad j = 0, \dots, n,$$

since $t = 0$ is a zero of B_k^n of order k. From this, by induction we find that $b_n = \cdots = b_0 = 0$. □

Definition 8.38 *The coefficients $b_0, \dots, b_n \in \mathbb{R}^m$ in the representation of a polynomial $p \in P_n^m$ through the Bernstein basis*

$$p(x) = \sum_{k=0}^{n} b_k B_k^n(x; a, b), \quad x \in [a,b], \tag{8.53}$$

are called control points, *or* Bézier points, *of p. The polygon determined by them is called the* Bézier polygon.

We now want to indicate that the graph of the polynomial p is closely related to the form of the Bézier polygon, and for this reason the graph of p is often referred to as *Bézier curve*. We first note that $p(a) = b_0$ and $p(b) = b_n$; i.e., both endpoints of the Bézier curve and the Bézier polygon coincide. Furthermore, from (8.49) it follows that the Bézier curve is contained in the *convex hull* $\text{con}\{b_0, \dots, b_n\}$ of the Bézier points. The convex hull

$$\text{con}\{b_0, \dots, b_n\} := \left\{ \sum_{k=0}^{n} \alpha_k b_k : \alpha_k \geq 0, \ \sum_{k=0}^{n} \alpha_k = 1 \right\}$$

is the smallest convex set containing the points $b_0, \dots, b_n$ (see Problem 8.19).

For computing the derivatives of a Bézier curve we first note that

$$\frac{d}{dt} B_k^n(t) = \binom{n}{k} [kt^{k-1}(1-t)^{n-k} - (n-k)t^k(1-t)^{n-k-1}]$$

implies that

$$(B_k^n)' = \begin{cases} -nB_0^{n-1}, & k = 0, \\ n(B_{k-1}^{n-1} - B_k^{n-1}), & k = 1, \dots, n-1, \\ nB_{n-1}^{n-1}, & k = n. \end{cases} \tag{8.54}$$

With this identity we are ready to establish the following theorem.

Theorem 8.39 *Let*

$$p(t) = \sum_{k=0}^{n} b_k B_k^n(t), \quad t \in [0,1],$$

be a Bézier polynomial on $[0,1]$. *Then*

$$p^{(j)}(t) = \frac{n!}{(n-j)!} \sum_{k=0}^{n-j} \Delta^j b_k B_k^{n-j}(t), \quad j = 1, \dots, n,$$

with the forward differences $\Delta^j b_k$ *recursively defined by*

$$\Delta^0 b_k := b_k, \quad \Delta^j b_k := \Delta^{j-1} b_{k+1} - \Delta^{j-1} b_k, \quad j = 1, \dots, n.$$

Proof. Obviously, the statement is true for $j = 0$. We assume that it has been proven for some $0 \le j < n$. Then with the aid of (8.54) we obtain

$$\begin{aligned} p^{(j+1)}(t) &= \frac{n!}{(n-j)!} \sum_{k=0}^{n-j} \Delta^j b_k \frac{d}{dt} B_k^{n-j}(t) \\ &= \frac{n!}{(n-j-1)!} \left\{ \sum_{k=1}^{n-j} \Delta^j b_k B_{k-1}^{n-j-1}(t) - \sum_{k=0}^{n-j-1} \Delta^j b_k B_k^{n-j-1}(t) \right\} \\ &= \frac{n!}{(n-j-1)!} \sum_{k=0}^{n-j-1} \left\{ \Delta^j b_{k+1} - \Delta^j b_k \right\} B_k^{n-j-1}(t) \\ &= \frac{n!}{[n-(j+1)]!} \sum_{k=0}^{n-(j+1)} \Delta^{j+1} b_k B_k^{n-(j+1)}(t), \end{aligned}$$

which establishes the assertion for $j + 1$. □

Corollary 8.40 *The polynomial from Theorem 8.39 has the derivatives*

$$p^{(j)}(0) = \frac{n!}{(n-j)!} \Delta^j b_0, \quad p^{(j)}(1) = \frac{n!}{(n-j)!} \Delta^j b_{n-j}$$

at the two endpoints.

From Corollary 8.40 we note that $p^{(j)}(0)$ depends only on $b_0, \ldots, b_j$ and that $p^{(j)}(1)$ depends only on $b_{n-j}, \ldots, b_n$. In particular, we have that

$$p'(0) = n(b_1 - b_0), \quad p'(1) = n(b_n - b_{n-1}); \tag{8.55}$$

i.e., at the two endpoints the Bézier curve has the same tangent lines as the Bézier polygon. Through the affine transformation (8.46) these results on the derivatives carry over to the general interval $[a, b]$.

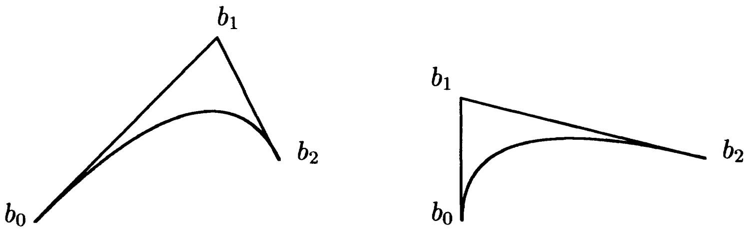


FIGURE 8.2. Bézier polynomials of degree two

Figure 8.2 illustrates by two Bézier polynomials of degree two in $\mathbb{R}^2$ how the shape of the curve is influenced by the location of the control points b_i. From (8.55) we also observe how to patch two Bézier polynomials of degree two together smoothly such that the tangent lines at the joints coincide, i.e., such that the two polynomials match up to a *Bézier spline* of degree two. The Bézier polynomials have the same tangent lines at the joints if the Bézier polygons do. This is illustrated by Figure 8.3.

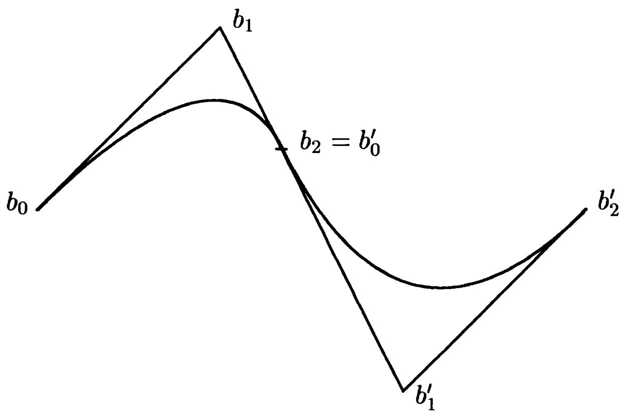


FIGURE 8.3. Bézier spline of degree two

We will conclude this section by describing the *de Casteljau algorithm* as a very stable and fast method for computing the function values $p(t)$ of

a Bézier polynomial. Given a Bézier polynomial

$$p(t) = \sum_{k=0}^{n} b_k B_k^n(t), \quad t \in [0,1],$$

we define the subpolynomials $b_i^k \in P_k^m$ by

$$b_i^k(t) := \sum_{j=0}^{k} b_{i+j} B_j^k(t) \tag{8.56}$$

for $i = 0, \ldots, n-k$ and $k = 0, \ldots, n$. For polynomials on $[a,b]$ we have an analogous definition for the subpolynomials. The subpolynomial b_i^k is a polynomial of degree k and has the $k+1$ control points $b_i, \ldots, b_{i+k}$. In particular, we have that $b_0^n = p$. Analogous to the Neville scheme of Theorem 8.9 we have the following recursion formula, which is the basis of the de Casteljau algorithm.

Theorem 8.41 *The subpolynomials b_i^k of a Bézier polynomial p of degree n satisfy the recursion formulae*

$$b_i^k(t) = (1-t) b_i^{k-1}(t) + t b_{i+1}^{k-1}(t) \tag{8.57}$$

for $i = 0, \ldots, n-k$ and $k = 1, \ldots, n$.

Proof. We insert the recursion formulae (8.51) and (8.52) for the Bernstein polynomials into the definition (8.56) for the subpolynomials and obtain

$$\begin{aligned} b_i^k(t) &= b_i B_0^k(t) + \sum_{j=1}^{k-1} b_{i+j} B_j^k(t) + b_{i+k} B_k^k(t) \\ &= \sum_{j=0}^{k-1} b_{i+j}(1-t) B_j^{k-1}(t) + \sum_{j=1}^{k} b_{i+j} t B_{j-1}^{k-1}(t) \\ &= (1-t) b_i^{k-1}(t) + t b_{i+1}^{k-1}(t), \end{aligned}$$

which establishes (8.57). □

Since $b_0^n(t) = p(t)$, starting the recursion with $b_k^0(t) = b_k$, from (8.57) we can compute $p(t)$ by successive convex combinations of the Bézier points $b_0, \ldots, b_n$, which clearly is a numerically stable procedure. Since (8.57) is similar in structure to the divided differences in Definition 8.4, the computations can be arranged in a tableau analogous to the one for the divided differences.

From the coefficients of the de Casteljau tableau we can construct two Bézier polynomials on the subintervals $[0,t]$ and $[t,1]$ that coincide with the original Bézier polynomial on the full interval $[0,1]$.

Theorem 8.42 *The Bézier polynomials*

$$p_1(x) := \sum_{k=0}^{n} b_0^k(t) B_k^n(x;0,t) \quad \text{and} \quad p_2(x) := \sum_{k=0}^{n} b_k^{n-k}(t) B_k^n(x;t,1)$$

with the coefficients b_0^k and b_k^{n-k} for $k = 0, \dots, n$ defined by the recursion (8.57) *satisfy*

$$p(x) = p_1(x) = p_2(x), \quad x \in \mathbb{R},$$

for arbitrary $0 < t < 1$.

Proof. Inserting the equivalent definition (8.56) of the subpolynomials and reordering the summation, we find that

$$p_1(x) = \sum_{k=0}^{n} \sum_{j=0}^{k} b_j B_j^k(t) B_k^n(x;0,t) = \sum_{j=0}^{n} b_j \sum_{k=j}^{n} B_j^k(t) B_k^n(x;0,t).$$

Hence the proof will be concluded by showing that

$$\sum_{k=j}^{n} B_j^k(t) B_k^n(x;0,t) = B_j^n(x), \quad x \in \mathbb{R}. \tag{8.58}$$

To establish this identity we make use of Definition 8.36 and obtain with the aid of the binomial formula that

$$\begin{aligned}
\sum_{k=j}^{n} B_j^k(t) B_k^n(x;0,t) &= \sum_{k=j}^{n} \binom{k}{j} (1-t)^{k-j} t^{j-n} \binom{n}{k} x^k (t-x)^{n-k} \\
&= \binom{n}{j} t^{j-n} \sum_{k=j}^{n} \binom{n-j}{k-j} (1-t)^{k-j} x^k (t-x)^{n-k} \\
&= \binom{n}{j} t^{j-n} x^j \sum_{k=0}^{n-j} \binom{n-j}{k} (1-t)^k x^k (t-x)^{n-j-k} \\
&= \binom{n}{j} x^j (1-x)^{n-j}.
\end{aligned}$$

Hence (8.58) is valid, and consequently $p_1 = p$. The proof of $p_2 = p$ is completely analogous, and it can also be obtained by a symmetry argument from $p_1 = p$. □

A natural choice in the subdivision of Theorem 8.42 is to break the interval in half by taking $t = 1/2$. Successively repeating the subdivision leads to a sequence of Bézier polygons that converges rapidly enough to the original Bézier curve to make this subdivision algorithm practical for an effective visualization of the curve on a computer.

Problems

8.1 Let $u_1, \ldots, u_n \in C[a,b]$ be linearly independent and let $x_1, \ldots, x_n \in [a,b]$ be distinct. For given values $y_1, \ldots, y_n \in \mathbb{R}$ consider the interpolation problem of finding a function $u \in U_n := \operatorname{span}\{u_1, \ldots, u_n\}$ with the property

$$u(x_j) = y_j, \quad j = 1, \ldots, n.$$

Show that the following three properties are equivalent:
(a) The interpolation problem is uniquely solvable for each given set of values $y_1, \ldots, y_n \in \mathbb{R}$.
(b) Each function $u \in U$ with zeros $u(x_j) = 0$ for $j = 1, \ldots, n$ vanishes identically.
(c) The $n \times n$ matrix with entries $u_k(x_j)$ for $j, k = 1, \ldots n$ is regular.

8.2 Consider the interpolation of $f(x) := x^4$ by a polynomial $p \in P_3$ with the four interpolation points $-1, 0, 1, 2$. Discuss the behavior of the error $p - f$ in the interval $[-1, 2]$.

8.3 Write a computer program for the Neville scheme of Theorem 8.9.

8.4 Show that the interpolation operator $L_n : C[a,b] \to P_n$ given by (8.7) is a linear operator. Show that it is a bounded operator if both the domain and range space are equipped with the maximum norm.

8.5 Let $x_0, \ldots, x_n \in \mathbb{R}$ be $n+1$ distinct points. Show that the *Vandermonde matrix* V with entries (x_j^k) for $j, k = 0, 1, \ldots n$ has determinant

$$\det V = \prod_{0 \le j < k \le n} (x_j - x_k).$$

8.6 Verify numerically the findings of Runge described in Example 8.14.

8.7 Verify the relations (8.12) for the Hermite factors.

8.8 Prove Theorem 8.19, i.e., the representation of the remainder in Hermite interpolation.

8.9 Given a twice continuously differentiable function $f : [a,b] \to \mathbb{R}$ and three points $x_0, x_1, x_2 \in [a,b]$ with $x_0 \neq x_2$, show that there exists a unique polynomial $p \in P_3$ for which

$$p(x_0) = f(x_0), \quad p'(x_1) = f(x_1), \quad p''(x_1) = f''(x_1), \quad p(x_2) = f(x_2).$$

Find a representation of the polynomial and give a representation of the remainder analogous to Theorem 8.10. (This is an example of *Hermite–Birkhoff interpolation*.)

8.10 *Inverse interpolation* can be used to solve nonlinear equations $f(x) = 0$ approximately by interchanging the roles of interpolation points and interpolation values. Find an approximation of the zero $x = 1.5$ for $f(x) = (4x+1)^3 - 343$ from the values of f at the four points $x = 0, 1, 2, 3$ by inverse cubic interpolation, i.e, by interpolating the inverse of f by a cubic polynomial with interpolation points $f(0), f(1), f(2), f(3)$ and interpolation values $0, 1, 2, 3$. For the computation use the Neville scheme. Are you satisfied with the accuracy of the result?

8.11 For the trigonometric interpolation from Theorem 8.24 with $2n+1$ equidistant interpolation points show that the Lagrange factors are given by

$$\ell_k(t) = F(t - t_k), \quad k = 0, \dots, 2n,$$

where

$$F(t) := \frac{1}{n+1} \, \frac{\sin\left(n + \dfrac{1}{2} + 1\right)t}{\sin\dfrac{t}{2}}$$

for $t \neq 0, \pm 2\pi, \pm 4\pi, \dots$. Prove that

$$\int_0^{2\pi} \ell_j(t)\ell_k(t)\,dt = \frac{2\pi}{2n+1}\,\delta_{jk}.$$

8.12 For the trigonometric interpolation from Theorem 8.24 with $2n+1$ equidistant interpolation points show that

$$\|L_n f - f\|_2 \to 0, \; n \to \infty,$$

for each continuous 2π-periodic function f.
Hint: With the aid of Problem 8.11, show that

$$\|L_n g\|_2 \leq \sqrt{2\pi}\,\|g\|_\infty$$

for all $n \in \mathbb{N}$ and all continuous 2π-periodic functions f and use the Weierstrass approximation theorem for periodic functions.

8.13 For the trigonometric interpolation from Theorem 8.24 with $2n+1$ equidistant interpolation points show that

$$\|L_n f - f\|_\infty \to 0, \; n \to \infty,$$

for each continuously differentiable 2π periodic function f.
Hint: For the functions $f_k(t) := e^{ikt}$ show that

$$\|L_n f_k - f_k\|_\infty \leq 2$$

for $n = 1, 2, \dots$ and $k = 0, \pm 1, \pm 2, \dots$, and use the fact that the Fourier series for continuously differentiable functions is uniformly convergent.

8.14 Write a computer program for the fast Fourier transform.

8.15 Given n distinct points $z_1, \dots, z_n \notin [a, b]$, n distinct points $x_1, \dots, x_n$ in $[a, b]$, and n values $y_1, \dots, y_n \in \mathbb{R}$, show that there exists a unique function of the form

$$u(x) = \sum_{k=1}^{n} \frac{a_k}{x_k + z_k}$$

with real coefficients $a_1, \dots, a_n$ such that

$$u(x_j) = y_j, \quad j = 1, \dots, n.$$

8.16 Verify the relations (8.34)–(8.36) for B-splines.

8.17 Use the fact that the second derivative of a cubic spline is a piecewise linear function to derive the linear system (8.43) without using the B-spline (8.36). Hint: On each subinterval integrate the piecewise linear function for s'' twice and eliminate the integration constants through the interpolation conditions. Then use the continuity of s' to obtain the linear system.

8.18 For the Bernstein polynomials show that

$$\sum_{k=0}^{n} \frac{k}{n} B_k^n(t) = t, \quad t \in \mathbb{R},$$

and

$$\sum_{k=0}^{n} \frac{k^2}{n^2} B_k^n(t) = \frac{n-1}{n} t^2 + \frac{t}{n}, \quad t \in \mathbb{R}.$$

8.19 Show that the convex hull

$$\operatorname{con}\{b_0, \dots, b_n\} := \left\{ \sum_{k=0}^{n} \alpha_k b_k : \alpha_k \geq 0, \ \sum_{k=0}^{n} \alpha_k = 1 \right\}$$

of $n+1$ points $b_0, \dots, b_n \in \mathbb{R}^m$ is convex and that $\operatorname{con}\{b_0, \dots, b_n\} \subset U$ for each convex set U with $b_0, \dots, b_n \in U$.

8.20 Give the Bézier representation of the (cubic) Hermite factors of Theorem 8.18 for the case of two interpolation points. Draw the graphs of the Hermite factors and their Bézier polygons.

9
Numerical Integration

Numerical integration formulae, or quadrature formulae, are methods for the approximate evaluation of definite integrals. They are needed for the computation of those integrals for which either the antiderivative of the integrand cannot be expressed in terms of elementary functions or for which the integrand is available only at discrete points, for example from experimental data. In addition and even more important, quadrature formulae provide a basic and important tool for the numerical solution of differential and integral equations, as we shall see in Chapters 10, 11, and 12.

The evaluation of planar areas bounded by curves is one of the oldest problems in science. Attempts to measure the area bounded by circles, ellipses, and parabolas were undertaken already by the Babylonians, Egyptians, and Greeks. However, a systematic analysis only became possible after the invention of calculus. Newton interpolated functions at equidistant points and integrated the interpolating polynomial and thus invented what now is known as the Newton–Cotes quadratures. Describing these interpolatory quadrature formulae will be the subject of Sections 9.1 and 9.2. Gauss was the first to notice that nonequidistant interpolation points lead, in general, to better accuracy for the resulting approximations to the integrals. In 1814 he presented a paper entitled "Methodus nova integralium valores per approximationem inveniendi" introducing quadrature formulae with the degree of accuracy considerably improved as compared with the Newton–Cotes formulae. These Gaussian quadrature formulae will be the subject of Section 9.3. The remaining part of this chapter is based on the Euler–Maclaurin expansion, which was found and published independently by Euler (1738) and Maclaurin (1737). We shall first employ the Euler–

Maclaurin expansion in our analysis of numerical integration of periodic functions. We will then use it to develop Romberg integration as a typical example for the use of the extrapolation method in order to increase the degree of accuracy. And finally, for integrands with endpoint singularities we will describe quadrature formulae that are based on a mesh that is graded towards the endpoints, and we will analyze the error with the help of the Euler–Maclaurin expansion.

For a comprehensive study of numerical integration methods including multidimensional integration we refer to [9, 17, 21, 57].

9.1 Interpolatory Quadratures

The most common quadrature formulae approximate the definite integral

$$Q(f) := \int_a^b f(x)\,dx \tag{9.1}$$

of a continuous function f over the interval $[a,b]$ with $a < b$ by a weighted sum

$$Q_n(f) := \sum_{k=0}^{n} a_k f(x_k) \tag{9.2}$$

with $n+1$ distinct *quadrature points* $x_0, \dots, x_n \in [a,b]$ and *quadrature weights* $a_0, \dots, a_n \in \mathbb{R}$. As one of the main applications of interpolation as developed in the previous chapter, an important group of quadrature formulae is obtained by integrating an interpolating polynomial instead of the integrand f, i.e., by approximating

$$\int_a^b f(x)\,dx \approx \int_a^b (L_n f)(x)\,dx,$$

where $L_n : C[a,b] \to P_n$ denotes the polynomial interpolation operator with interpolation points $x_0, \dots, x_n$ introduced in Section 8.1 (see (8.7)). Note that both the integral Q and the quadrature formula Q_n represent linear operators from $C[a,b]$ into $\mathbb{R}$.

Theorem 9.1 *The* polynomial interpolatory quadrature *of order n defined by*

$$Q_n(f) := \int_a^b (L_n f)(x)\,dx \tag{9.3}$$

is of the form (9.2) *with the weights given by*

$$a_k = \frac{1}{q'_{n+1}(x_k)} \int_a^b \frac{q_{n+1}(x)}{x - x_k}\,dx, \quad k = 0, \dots, n, \tag{9.4}$$

where $q_{n+1}(x) := (x - x_0) \cdots (x - x_n)$.

Proof. From (8.2) we obtain

$$\int_a^b (L_n f)(x)\,dx = \sum_{k=0}^{n} f(x_k) \int_a^b \ell_k(x)\,dx$$

with

$$a_k = \int_a^b \ell_k(x)\,dx = \int_a^b \prod_{\substack{j=0\\j\neq k}}^{n} \frac{x - x_j}{x_k - x_j}\,dx,$$

whence (9.4) follows by rewriting the product. □

The following theorem describes an equivalent definition of polynomial interpolatory quadratures.

Theorem 9.2 *Given $n+1$ distinct quadrature points $x_0, \ldots, x_n \in [a, b]$, the interpolatory quadrature* (9.3) *of order n is uniquely determined by its property of integrating all polynomials $p \in P_n$ exactly, i.e., by the property*

$$\sum_{k=0}^{n} a_k p(x_k) = \int_a^b p(x)\,dx \tag{9.5}$$

for all $p \in P_n$.

Proof. From (9.3) and $L_n p = p$ for all $p \in P_n$ it follows that

$$\sum_{k=0}^{n} a_k p(x_k) = \int_a^b (L_n p)(x)\,dx = \int_a^b p(x)\,dx;$$

i.e., the quadrature is exact for all $p \in P_n$. On the other hand, from (9.5) we obtain

$$\sum_{k=0}^{n} a_k f(x_k) = \sum_{k=0}^{n} a_k (L_n f)(x_k) = \int_a^b (L_n f)(x)\,dx$$

for all $f \in C[a, b]$; i.e., the quadrature is an interpolatory quadrature. □

Theorem 9.3 *The polynomial interpolatory quadrature of order n with equidistant quadrature points*

$$x_k = a + kh, \quad k = 0, \ldots, n,$$

and step width $h = (b-a)/n$ is called the Newton–Cotes quadrature formula *of order n. Its weights are given by*

$$a_k = h\,\frac{(-1)^{n-k}}{k!\,(n-k)!} \int_0^n \prod_{\substack{j=0\\j\neq k}}^{n} (z - j)\,dz, \quad k = 0, \ldots, n, \tag{9.6}$$

and have the symmetry property $a_k = a_{n-k}$, $k = 0, \ldots, n$.

Proof. The weights are obtained from (9.4) by substituting $x = x_0 + hz$ and observing that

$$q_{n+1}(x) = h^{n+1} \prod_{j=0}^{n} (z - j)$$

and

$$q'_{n+1}(x_k) = (-1)^{n-k} k!\,(n-k)!\,h^n.$$

The symmetry $a_k = a_{n-k}$ follows by substituting $z = n - y$. □

These quadrature formulae were first discovered by Newton and also carry the name of Cotes because of his systematic account of Newton's integration rules in 1711. The Newton–Cotes quadrature formula of order $n = 1$ is known as the *trapezoidal rule.* Its weights can be obtained either from evaluating (9.6) or more easily from the exactness conditions of Theorem 9.2. For the interval $[-1, 1]$, these conditions are given by

$$a_0 + a_1 = \int_{-1}^{1} dx = 2,$$

$$-a_0 + a_1 = \int_{-1}^{1} x\,dx = 0,$$

and imply that $a_0 = a_1 = 1$. Hence, for a general interval the trapezoidal rule has the form

$$\int_a^b f(x)\,dx \approx \frac{b-a}{2}\,[f(a) + f(b)] = \frac{h}{2}\,[f(x_0) + f(x_1)].$$

Geometrically speaking, the trapezoidal rule approximates the integral of f by the integral of the straight line connecting the two points $(a, f(a))$ and $(b, f(b))$. Hence, the approximate value coincides with the area of the trapezoid with the four corners $(a, 0)$, $(b, 0)$, $(a, f(a))$, and $(b, f(b))$.

The Newton–Cotes quadrature formula of order $n = 2$ was already known to Kepler in 1612 and Cavalieri in 1639 and is called *Simpson's rule,* since Simpson rediscovered it in 1743. Its weights are obtained from the exactness conditions

$$a_0 + a_1 + a_2 = \int_{-1}^{1} dx = 2,$$

$$-a_0 \qquad + a_2 = \int_{-1}^{1} x\,dx = 0,$$

$$a_0 \qquad + a_2 = \int_{-1}^{1} x^2 dx = \frac{2}{3}\,,$$

which imply that $a_0 = a_2 = 1/3$ and $a_1 = 4/3$. Hence, for a general interval, Simpson's rule is given by

$$\int_a^b f(x)\,dx \approx \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] = \frac{h}{3}\,[f(x_0)+4f(x_1)+f(x_2)].$$

Geometrically speaking, Simpson's rule approximates the integral of f by the integral of the parabola through the three points $(a, f(a))$, $(\frac{a+b}{2}, f(\frac{a+b}{2}))$, and $(b, f(b))$.

Table 9.1 gives the weights of the first four Newton–Cotes formulae (with the common factor $h = (b-a)/n$ omitted).

TABLE 9.1. Weights of Newton–Cotes formulae

n	a_k					
1	$\frac{1}{2}$	$\frac{1}{2}$				Trapezoidal rule
2	$\frac{1}{3}$	$\frac{4}{3}$	$\frac{1}{3}$			Simpson's rule
3	$\frac{3}{8}$	$\frac{9}{8}$	$\frac{9}{8}$	$\frac{3}{8}$		Newton's three-eights rule
4	$\frac{14}{45}$	$\frac{64}{45}$	$\frac{24}{45}$	$\frac{64}{45}$	$\frac{14}{45}$	Milne's rule

For $n \geq 8$ some of the weights of the Newton–Cotes formulae become negative (see Problem 9.4). Since this might lead to negative approximations for integrals with positive integrands, the higher-order Newton–Cotes rules cannot be recommended for numerical purposes.

We will carry out the error analysis for the Newton–Cotes formulae only for the two most important cases, $n = 1$ and $n = 2$, i.e., the trapezoidal rule and Simpson's rule.

Theorem 9.4 *Let $f : C[a, b] \to \mathbb{R}$ be twice continuously differentiable. Then the error for the trapezoidal rule can be represented in the form*

$$\int_a^b f(x)\,dx - \frac{b-a}{2}\,[f(a) + f(b)] = -\frac{h^3}{12}\,f''(\xi) \tag{9.7}$$

with some $\xi \in [a, b]$ and $h = b - a$.

Proof. Let $L_1 f$ denote the linear interpolation of f at the interpolation points $x_0 = a$ and $x_1 = b$. By construction of the trapezoidal rule we have

that the error

$$E_1(f) := \int_a^b f(x)\,dx - \frac{b-a}{2}\,[f(a) + f(b)]$$

is given by

$$E_1(f) = \int_a^b [f(x) - (L_1 f)(x)]\,dx = \int_a^b (x-a)(x-b)\,\frac{f(x) - (L_1 f)(x)}{(x-a)(x-b)}\,dx.$$

Since the first factor of the integrand is nonpositive on $[a, b]$ and since by l'Hôpital's rule the second factor is continuous, from the mean value theorem for integrals we obtain that

$$E_1(f) = \frac{f(z) - (L_1 f)(z)}{(z-a)(z-b)} \int_a^b (x-a)(x-b)\,dx$$

for some $z \in [a, b]$. From this, with the aid of the error representation for linear interpolation from Theorem 8.10 and the integral

$$\int_a^b (x-a)(x-b)\,dx = -\frac{h^3}{6}\,,$$

the assertion of the theorem follows. □

We explicitly note that (9.7) cannot be obtained by integrating the interpolation error representation (8.8), since we do not know whether the intermediate point ξ in (8.8) depends continuously on x.

By construction, Simpson's rule integrates polynomials of degree less than or equal to two exactly. In addition, it also integrates polynomials of degree three exactly. By linearity, to show this it suffices to prove it for one polynomial of degree three. For the polynomial

$$q_3(x) = (x - x_0)(x - x_1)(x - x_2)$$

both the integral and the value obtained from Simpson's rule are zero. Hence, this polynomial of degree three is integrated exactly by Simpson's rule.

Theorem 9.5 *Let $f : C[a, b] \to \mathbb{R}$ be four-times continuously differentiable. Then the error for Simpson's rule can be represented in the form*

$$\int_a^b f(x)\,dx - \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right] = -\frac{h^5}{90}\,f^{(4)}(\xi) \qquad (9.8)$$

for some $\xi \in [a, b]$ and $h = (b-a)/2$.

Proof. Let $L_2 f$ denote the quadratic interpolation polynomial for f at the interpolation points $x_0 = a$, $x_1 = (a+b)/2$, and $x_2 = b$. By construction of Simpson's rule we have that the error

$$E_2(f) := \int_a^b f(x)\,dx - \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right]$$

is given by

$$E_2(f) = \int_a^b [f(x) - (L_2 f)(x)]\,dx. \tag{9.9}$$

Consider the cubic polynomial

$$p(x) := (L_2 f)(x) + \frac{4}{(b-a)^2}\,[(L_2 f)'(x_1) - f'(x_1)]\,q_3(x), \tag{9.10}$$

where $q_3(x) = (x - x_0)(x - x_1)(x - x_2)$. Obviously, p has the interpolation properties

$$p(x_k) = f(x_k), \quad k = 0, 1, 2, \quad \text{and} \quad p'(x_1) = f'(x_1).$$

Since $\int_a^b q_3(x)\,dx = 0$, from (9.9) and (9.10) we can conclude that

$$E_2(f) = \int_a^b [f(x) - p(x)]\,dx,$$

and consequently

$$E_2(f) = \int_a^b (x - x_0)(x - x_1)^2(x - x_2)\,\frac{f(x) - p(x)}{(x - x_0)(x - x_1)^2(x - x_2)}\,dx.$$

As in the proof of Theorem 9.4, the first factor of the integrand is nonpositive on $[a, b]$, and the second factor is continuous. Hence, by the mean value theorem for integrals, we obtain that

$$E_2(f) = \frac{f(z) - p(z)}{(z - x_0)(z - x_1)^2(z - x_2)}\int_a^b (x - x_0)(x - x_1)^2(x - x_2)\,dx$$

for some $z \in [a, b]$. Analogous to Theorem 8.10, it can be shown that

$$f(z) - p(z) = \frac{f^{(4)}(\xi)}{4!}\,(z - x_0)(z - x_1)^2(z - x_2)$$

for some $\xi \in [a, b]$. From this, with the aid of the integral

$$\int_a^b (x - x_0)(x - x_1)^2(x - x_2)\,dx = -\frac{(b-a)^5}{120}\,,$$

we conclude the statement of the theorem. □

Example 9.6 The approximation of

$$\ln 2 = \int_0^1 \frac{dx}{1+x}$$

by the trapezoidal rule yields

$$\ln 2 \approx \frac{1}{2}\left[1 + \frac{1}{2}\right] = 0.75.$$

For $f(x) := 1/(1+x)$ we have

$$\frac{h^3}{12}\,\|f''\|_\infty = \frac{1}{6}\,,$$

and hence, from Theorem 9.4, we obtain the estimate $|\ln 2 - 0.75| \le 0.167$ as compared to the true error $\ln 2 - 0.75 = -0.056\ldots$.

Simpson's rule yields

$$\ln 2 \approx \frac{1}{6}\left[1 + \frac{4}{1+\frac{1}{2}} + \frac{1}{2}\right] = \frac{25}{36} = 0.6944\ldots,$$

and from Theorem 9.5 and

$$\frac{h^5}{90}\,\|f^{(4)}\|_\infty = \frac{1}{120}$$

we find the estimate $|\ln 2 - 0.6944| \le 0.0084$ as compared to the true error $\ln 2 - 25/36 = -0.0012\ldots$. □

In order to increase the accuracy, instead of using higher order Newton–Cotes rules it is more practical to use so-called *composite rules.* These are obtained by subdividing the interval of integration and then applying a fixed rule with low interpolation order to each of the subintervals. The most frequently used quadrature rules of this type are the *composite trapezoidal rule* and the *composite Simpson's rule.*

Let $x_k = a + kh$, $k = 0, \ldots, n$, be an equidistant subdivision with step size $h = (b-a)/n$. Then the composite trapezoidal rule is given by

$$T_h(f) := h\left[\frac{1}{2}\,f(x_0) + f(x_1) + \cdots + f(x_{n-1}) + \frac{1}{2}\,f(x_n)\right]$$

for $f \in C[a,b]$.

Theorem 9.7 *Let $f : [a,b] \to \mathbb{R}$ be twice continuously differentiable. Then the error for the composite trapezoidal rule is given by*

$$\int_a^b f(x)\,dx - T_h(f) = -\frac{b-a}{12}\,h^2 f''(\xi)$$

for some $\xi \in [a,b]$.

Proof. By Theorem 9.4 we have that

$$\int_a^b f(x)\,dx - T_h(f) = -\frac{h^3}{12}\sum_{k=1}^{n} f''(\xi_k),$$

where $a \le \xi_1 \le \xi_2 \le \cdots \le \xi_n \le b$. From

$$n \min_{x\in[a,b]} f''(x) \le \sum_{k=1}^{n} f''(\xi_k) \le n \max_{x\in[a,b]} f''(x)$$

and the continuity of f'' we conclude that there exists $\xi \in [a,b]$ such that

$$\sum_{k=1}^{n} f''(\xi_k) = nf''(\xi),$$

and the proof is finished. □

Let n be even. Then the composite Simpson's rule is given by

$$S_h(f) := \frac{h}{3}\,[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \cdots + 2f(x_{n-2}) + 4f(x_{n-1}) + f(x_n)]$$

for $f \in C[a,b]$. Its error can be represented and estimated as follows.

Theorem 9.8 *Let $f : [a,b] \to \mathbb{R}$ be four-times continuously differentiable. Then the error for the composite Simpson's rule is given by*

$$\int_a^b f(x)\,dx - S_h(f) = -\frac{b-a}{180}\,h^4 f^{(4)}(\xi)$$

for some $\xi \in [a,b]$.

Proof. Using Theorem 9.5, the proof is analogous to the proof of Theorem 9.7. □

Table 9.2 gives the error between the exact value of the integral from Example 9.6 and its numerical approximation by the composite trapezoidal rule and the composite Simpson's rule. Clearly, if the number n of quadrature points is doubled, i.e., if the step size h is halved, then the error for the trapezoidal rule is reduced by the factor $1/4$ and for Simpson's rule by the factor $1/16$, as predicted in Theorems 9.7 and 9.8.

TABLE 9.2. Trapezoidal and Simpson's rule for Example 9.6

n	Trapezoidal rule	Simpson's rule
1	−0.05685282	
2	−0.01518615	−0.00129726
4	−0.00387663	−0.00010679
8	−0.00097467	−0.00000735
16	−0.00024402	−0.00000047
32	−0.00006103	−0.00000003

9.2 Convergence of Quadrature Formulae

Definition 9.9 *A sequence (Q_n) of quadrature formulae is called convergent if*

$$Q_n(f) \to Q(f) = \int_a^b f(x)\,dx, \quad n \to \infty,$$

for all $f \in C[a,b]$.

Theorem 9.10 (Szegő) *Let*

$$Q_n(f) = \sum_{k=0}^{n} a_k^{(n)} f(x_k^{(n)})$$

be a sequence of quadrature formulae that converges for all polynomials, i.e,

$$\lim_{n\to\infty} Q_n(p) = Q(p) \tag{9.11}$$

for all polynomials p, and that is uniformly bounded, i.e., there exists a constant $C > 0$ such that

$$\sum_{k=0}^{n} |a_k^{(n)}| \le C \tag{9.12}$$

for all $n \in \mathbb{N}$. Then the sequence (Q_n) is convergent.

Proof. Let $f \in C[a,b]$ and $\varepsilon > 0$ be arbitrary. By the Weierstrass approximation theorem (see [16]) there exists a polynomial p such that

$$\|f - p\|_\infty \le \frac{\varepsilon}{2(C + b - a)}\,.$$

Then, since by (9.11) we have $Q_n(p) \to Q(p)$ as $n \to \infty$, there exists $N(\varepsilon) \in \mathbb{N}$ such that

$$|Q_n(p) - Q(p)| \le \frac{\varepsilon}{2}$$

for all $n \geq N(\varepsilon)$. Now with the aid of the triangle inequality and using (9.12) we can estimate

$$|Q_n(f) - Q(f)| \leq \sum_{k=0}^{n} |a_k^{(n)}| \, |f(x_k^{(n)}) - p(x_k^{(n)})| + |Q_n(p) - Q(p)|$$

$$+ \int_a^b |p(x) - f(x)| \, dx$$

$$\leq \frac{C\varepsilon}{2(C+b-a)} + \frac{\varepsilon}{2} + \frac{(b-a)\varepsilon}{2(C+b-a)} = \varepsilon$$

for all $N \geq N(\varepsilon)$; i.e., $Q_n(f) \to Q(f)$ for $n \to \infty$. □

A quadrature formula

$$Q_n(f) = \sum_{k=0}^{n} a_k f(x_k)$$

defines a bounded linear operator $Q_n : C[a,b] \to \mathbb{R}$ with the norm given by

$$\|Q_n\|_\infty = \sum_{k=0}^{n} |a_k|. \tag{9.13}$$

To prove this, we note the estimate

$$|Q_n f| \leq \|f\|_\infty \sum_{k=0}^{n} |a_k|,$$

which implies that Q_n is a bounded operator and that the operator norm is less than or equal to the right-hand side of (9.13). Equality in (9.13) follows by choosing f to be a continuous piecewise linear function with $\|f\|_\infty = 1$ and $f(x_k)a_k = |a_k|$ for $k = 0, \ldots, n$. From (9.13) and the uniform boundedness principle, Theorem 12.7, it can be seen that the two conditions of Theorem 9.10 are also necessary for convergence of a sequence of quadrature formulae.

Corollary 9.11 (Steklov) *Assume that the sequence (Q_n) of quadrature formulae converges for all polynomials and that all the weights are nonnegative. Then the sequence (Q_n) is convergent.*

Proof. This follows from

$$\sum_{k=0}^{n} |a_k^{(n)}| = \sum_{k=0}^{n} a_k^{(n)} = Q_n(1) \to \int_a^b dx = b - a, \quad n \to \infty,$$

and the preceding Theorem 9.10. □

From Theorems 9.7 and 9.8 and Corollary 9.11 we observe that the composite trapezoidal rule and the composite Simpson's rule are convergent. On the other hand, using the fact that the conditions of Theorem 9.10 are necessary for convergence, it can be shown that the Newton–Cotes quadratures do not converge for all continuous functions (see Problem 9.5).

9.3 Gaussian Quadrature Formulae

Given the arbitrary quadrature points $x_0, \dots, x_n$ in $[a, b]$, the quadrature weights $a_0, \dots, a_n$ of a polynomial interpolatory quadrature are determined such that all polynomials of degree less than or equal to n are integrated exactly. In this section we will examine the problem of whether the quadrature points can be chosen in such a way that polynomials of degree less than or equal to $2n + 1$ are also integrated exactly. Obviously, to achieve this degree of exactness the quadrature points and the quadrature weights have to satisfy the conditions

$$\sum_{k=0}^{n} a_k x_k^i = \int_a^b x^i dx, \quad i = 0, \dots, 2n+1.$$

We shall see that this system of $2n + 2$ nonlinear equations for the $2n + 2$ unknowns $x_0, \dots, x_n \in [a, b]$ and $a_0, \dots, a_n \in \mathbb{R}$ has a unique solution and that for this solution the points $x_0, \dots, x_n$ are distinct.

We shall proceed slightly more generally by considering quadrature formulae for the integral

$$Q(f) := \int_a^b w(x) f(x)\, dx, \tag{9.14}$$

where w denotes some *weight function.* We assume that $w : (a, b) \to \mathbb{R}$ is continuous and positive and that the integral $\int_a^b w(x)\, dx$ exists. Typical examples are given by

$$w(x) = 1, \quad w(x) = \sqrt{1 - x^2}, \quad w(x) = \frac{1}{\sqrt{1 - x^2}},$$

where for the two latter cases the interval is assumed to be $[a, b] = [-1, 1]$. Analogously to the case $w(x) = 1$, interpolatory quadrature rules for (9.14) are obtained by replacing f through its interpolation polynomial $L_n f$ and then integrating exactly, i.e., by approximating Qf through

$$Q_n(f) := \int_a^b w(x)(L_n f)(x)\, dx.$$

Note that the separation of a weight function w for interpolatory quadrature formulae has the advantage that in general, $wL_n f$ is a better approximation to wf than $L_n(wf)$ due to possible singularities of w and its derivatives at the endpoints of the interval.

Definition 9.12 *A quadrature formula*

$$\int_a^b w(x)f(x)\,dx \approx \sum_{k=0}^n a_k f(x_k)$$

with $n+1$ distinct quadrature points is called a Gaussian quadrature formula *if it integrates all polynomials $p \in P_{2n+1}$ exactly, i.e., if*

$$\sum_{k=0}^n a_k p(x_k) = \int_a^b w(x)p(x)\,dx \tag{9.15}$$

for all $p \in P_{2n+1}$.

Lemma 9.13 *Let $x_0, \ldots, x_n$ be the $n+1$ distinct quadrature points of a Gaussian quadrature formula. Then*

$$\int_a^b w(x)q_{n+1}(x)q(x)\,dx = 0 \tag{9.16}$$

for $q_{n+1}(x) := (x-x_0)\cdots(x-x_n)$ and all $q \in P_n$.

Proof. Since $q_{n+1}q \in P_{2n+1}$ and $q_{n+1}(x_k) = 0$, we have that

$$\int_a^b w(x)q_{n+1}(x)q(x)\,dx = \sum_{k=0}^n a_k q_{n+1}(x_k)q(x_k) = 0$$

for all $q \in P_n$. □

Lemma 9.14 *Let $x_0, \ldots, x_n$ be $n+1$ distinct points satisfying the condition* (9.16)*. Then the corresponding polynomial interpolatory quadrature is a Gaussian quadrature formula.*

Proof. Let L_n denote the polynomial interpolation operator for the interpolation points $x_0, \ldots, x_n$. By construction, for the interpolatory quadrature we have

$$\sum_{k=0}^n a_k f(x_k) = \int_a^b w(x)(L_n f)(x)\,dx \tag{9.17}$$

for all $f \in C[a,b]$. Each $p \in P_{2n+1}$ can be represented in the form

$$p = L_n p + q_{n+1}q$$

for some $q \in P_n$, since the polynomial $p - L_n p$ vanishes at the points $x_0, \ldots, x_n$. Then from (9.16) and (9.17) we obtain that

$$\int_a^b w(x)p(x)\,dx = \int_a^b w(x)(L_n p)(x)\,dx = \sum_{k=0}^n a_k p(x_k)$$

for all $p \in P_{2n+1}$. □

Lemma 9.15 *There exists a unique sequence (q_n) of polynomials of the form $q_0 = 1$ and*

$$q_n(x) = x^n + r_{n-1}(x), \quad n = 1, 2, \ldots,$$

with $r_{n-1} \in P_{n-1}$ satisfying the orthogonality *relation*

$$\int_a^b w(x) q_n(x) q_m(x) dx = 0, \quad n \neq m, \tag{9.18}$$

and

$$P_n = \operatorname{span}\{q_0, \ldots, q_n\}, \quad n = 0, 1, \ldots. \tag{9.19}$$

Proof. This follows by the Gram–Schmidt orthogonalization procedure from Theorem 3.18 applied to the linearly independent functions $u_n(x) := x^n$ for $n = 0, 1, \ldots$ and the scalar product

$$(f, g) := \int_a^b w(x) f(x) g(x)\, dx$$

for $f, g \in C[a, b]$. The positive definiteness of the scalar product is a consequence of w being positive in (a, b). □

Lemma 9.16 *Each of the orthogonal polynomials q_n from Lemma 9.15 has n simple zeros in (a, b).*

Proof. For $m = 0$, from (9.18) we have that

$$\int_a^b w(x) q_n(x)\, dx = 0$$

for $n > 0$. Hence, since w is positive on (a, b), the polynomial q_n must have at least one zero in (a, b) where the sign of q_n changes. Denote by $x_1, \ldots, x_m$ the zeros of q_n in (a, b) where q_n changes its sign. We assume that $m < n$ and set $r_m(x) := (x - x_1) \cdots (x - x_m)$. Then $r_m \in P_{n-1}$ and therefore

$$\int_a^b w(x) r_m(x) q_n(x)\, dx = 0.$$

However, this integral must be different from zero, since $r_m q_n$ does not change its sign on (a, b) and does not vanish identically. Hence, we have arrived at a contradiction, and consequently $m = n$. □

Theorem 9.17 *For each $n = 0, 1, \ldots$ there exists a unique Gaussian quadrature formula of order n. Its quadrature points are given by the zeros of the orthogonal polynomial q_{n+1} of degree $n + 1$.*

Proof. This is a consequence of Lemmas 9.13–9.16. □

Theorem 9.18 *The weights of the Gaussian quadrature formulae are all positive.*

Proof. Define

$$f_k(x) := \left[\frac{q_{n+1}(x)}{x - x_k}\right]^2, \quad k = 0, \dots, n.$$

Then

$$a_k[q'_{n+1}(x_k)]^2 = \sum_{j=0}^{n} a_j f_k(x_j) = \int_a^b w(x) f_k(x)\, dx > 0,$$

since $f_k \in P_{2n}$, and the theorem is proven. □

Corollary 9.19 *The sequence of Gaussian quadrature formulae is convergent.*

Proof. For each polynomial p we have

$$Q_n(p) = \int_a^b w(x) p(x)\, dx,$$

provided that $2n + 1$ is greater than or equal to the degree of p. From their proofs it is obvious that Theorem 9.10 and its Corollary 9.11 remain valid for the integral with the weight function w. Hence, the statement of the theorem follows from Theorem 9.18. □

Theorem 9.20 *Let $f \in C^{2n+2}[a,b]$. Then the error for the Gaussian quadrature formula of order n is given by*

$$\int_a^b w(x) f(x)\, dx - \sum_{k=0}^{n} a_k f(x_k) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_a^b w(x)[q_{n+1}(x)]^2 dx$$

for some $\xi \in [a,b]$.

Proof. Recall the Hermite interpolation polynomial $H_n f \in P_{2n+1}$ for f from Theorem 8.18. Since $(H_n f)(x_k) = f(x_k)$, $k = 0, \dots, n$, for the error

$$E_n(f) := \int_a^b w(x) f(x)\, dx - \sum_{k=0}^{n} a_k f(x_k)$$

we can write

$$E_n(f) = \int_a^b w(x)[f(x) - (H_n f)(x)]\, dx.$$

Then as in the proofs of Theorems 9.7 and 9.8, using the mean value theorem we obtain

$$E_n(f) = \frac{f(z) - (H_n f)(z)}{[q_{n+1}(z)]^2} \int_a^b w(x)[q_{n+1}(x)]^2 dx$$

for some $z \in [a,b]$. Now the proof is finished with the aid of the error representation for Hermite interpolation from Theorem 8.19. □

Example 9.21 We consider the Gaussian quadrature formulae for the weight function

$$w(x) = \frac{1}{\sqrt{1-x^2}}\,, \quad x \in [-1,1].$$

The *Chebyshev polynomial* T_n of degree n is defined by

$$T_n(x) := \cos(n \arccos x), \quad -1 \le x \le 1.$$

Obviously $T_0(x) = 1$ and $T_1(x) = x$. From the addition theorem for the cosine function, $\cos(n+1)t + \cos(n-1)t = 2\cos t \cos nt$, we can deduce the recursion formula

$$T_{n+1}(x) + T_{n-1}(x) = 2xT_n(x), \quad n = 1, 2, \ldots.$$

Hence we have that $T_n \in P_n$ with leading term

$$T_n(x) = 2^{n-1}x^n + \cdots, \quad n = 1, 2, \ldots.$$

Substituting $x = \cos t$ we find that

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}\,dx = \int_0^\pi \cos nt \cos mt\,dt = \begin{cases} \pi, & n = m = 0, \\ \frac{\pi}{2}, & n = m > 0, \\ 0, & n \ne m. \end{cases}$$

Hence, the orthogonal polynomials q_n of Lemma 9.15 are given by $q_n = 2^{1-n}T_n$. The zeros of T_n and hence the quadrature points are given by

$$x_k = \cos\left(\frac{2k+1}{2n}\,\pi\right), \quad k = 0, \ldots, n-1.$$

The weights can be most easily derived from the exactness conditions

$$\sum_{k=0}^{n-1} a_k\, T_m(x_k) = \int_{-1}^{1} \frac{T_m(x)}{\sqrt{1-x^2}}\,dx, \quad m = 0, \ldots, n-1,$$

for the interpolation quadrature, i.e., from

$$\sum_{k=0}^{n-1} a_k \cos\frac{(2k+1)m}{2n}\,\pi = \begin{cases} \pi, & m = 0, \\ 0, & m = 1, \ldots, n-1. \end{cases}$$

From our analysis of trigonometric interpolation, i.e., from (8.19), we see that the unique solution of this linear system is given by

$$a_k = \frac{\pi}{n}\,, \quad k = 0, \ldots, n-1.$$

Hence, for $n = 1, 2, \dots$ the *Gauss–Chebyshev quadrature* of order $n - 1$ is given by

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\, dx \approx \frac{\pi}{n} \sum_{k=0}^{n-1} f\left(\cos \frac{2k+1}{2n}\, \pi\right).$$

From Theorem 9.20 we have the error representation

$$\int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\, dx - \frac{\pi}{n} \sum_{k=0}^{n-1} f\left(\cos \frac{2k+1}{2n}\, \pi\right) = \frac{\pi f^{(2n)}(\xi)}{2^{2n-1}(2n)!}$$

for some $\xi \in [-1, 1]$. □

Example 9.22 We now consider the weight function

$$w(x) = 1, \quad x \in [-1, 1].$$

The *Legendre polynomial* L_n of degree n is defined by

$$L_n(x) := \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n.$$

Obviously, $L_n \in P_n$. If $m < n$, by repeated partial integration we see that

$$\int_{-1}^{1} x^m \frac{d^n}{dx^n} (x^2 - 1)^n dx = 0,$$

since $(x^2 - 1)^n$ has zeros of order n at the endpoints -1 and 1. Therefore,

$$\int_{-1}^{1} L_n(x) L_m(x)\, dx = 0, \quad n \neq m.$$

The zeros of the Legendre polynomials, and therefore the quadrature points and weights of the corresponding *Gauss–Legendre quadratures*, cannot be given explicitly by a simple expression. We consider only the cases $n = 1$ and $n = 2$ and note that

$$q_0(x) = 1, \quad q_1(x) = x, \quad q_2(x) = x^2 - \frac{1}{3},$$

where the coefficient of q_2 can be determined from $\int_{-1}^{1} q_2(x)\, dx = 0$.

The quadrature point for the first Gauss–Legendre formula is $x_1 = 0$, and the weight a_1 can be obtained from the exactness condition

$$a_1 = \int_{-1}^{1} dx = 2.$$

Hence the first Gauss–Legendre formula is given by

$$\int_{-1}^{1} f(x)\, dx \approx 2 f(0) \tag{9.20}$$

with the error representation

$$\int_{-1}^{1} f(x)\,dx - 2f(0) = \frac{1}{3}\,f''(\xi)$$

for some $\xi \in [-1,1]$. The coefficient of the derivative on the right-hand side follows most easily by inserting $f(x) = x^2$. For obvious reasons, this Gauss–Legendre formula is also known as the *midpoint rule.*

The quadrature points for the second Gauss–Legendre formula are $x_1 = -1/\sqrt{3}$ and $x_2 = 1/\sqrt{3}$. The weights can be obtained from the exactness conditions

$$a_1 + a_2 = \int_{-1}^{1} dx = 2,$$

$$a_1x_1 + a_2x_2 = \int_{-1}^{1} x\,dx = 0,$$

and they have the values $a_1 = a_2 = 1$. Hence the second Gauss–Legendre formula is given by

$$\int_{-1}^{1} f(x)\,dx \approx f\left(\frac{-1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right)$$

with the error representation

$$\int_{-1}^{1} f(x)\,dx - f\left(\frac{-1}{\sqrt{3}}\right) - f\left(\frac{1}{\sqrt{3}}\right) = \frac{1}{135}\,f^{(4)}(\xi)$$

for some $\xi \in [-1,1]$. The coefficient on the right-hand side follows by inserting $f(x) = x^4$. □

From the Gaussian quadrature formula

$$\int_{-1}^{1} g(z)\,dz \approx \sum_{k=0}^{n} a_k g(x_k)$$

of order n for the interval $[-1,1]$, by substituting

$$x = \frac{a+b}{2} + \frac{b-a}{2}\,z$$

and $f(x) = g(z)$ we obtain the Gaussian quadrature formula

$$\int_{a}^{b} f(x)\,dx \approx \frac{b-a}{2}\sum_{k=0}^{n} a_k f\left(\frac{a+b}{2} + \frac{b-a}{2}\,x_k\right)$$

for an arbitrary interval $[a,b]$. The error representation

$$\int_{-1}^{1} g(z)\,dz - \sum_{k=0}^{n} a_k g(x_k) = \frac{g^{(2n)}(\zeta)}{(2n+2)!}\int_{-1}^{1} [q_{n+1}(x)]^2\,dx$$

with $\zeta \in [-1, 1]$ can be transformed accordingly. Subdividing the interval $[a, b]$ into m equidistant subintervals with step width $h = (b - a)/m$ and then applying to each subinterval the Gaussian quadrature formula of order n, we obtain the *composite Gaussian quadrature*

$$\int_a^b f(x)\,dx \approx \frac{h}{2} \sum_{j=0}^{m-1} \sum_{k=0}^{n} a_k\, f\left(a + jh + \frac{h}{2} + \frac{h}{2}\, x_k\right)$$

with an error of order $O(h^{2n})$. These composite Gaussian rules are used quite frequently in practice. We illustrate their convergence behavior by Table 9.3, which gives the error between the exact value of the integral from Example 9.6 and its numerical approximation by composite Gaussian quadrature of orders one and two. As predicted by our error analysis, if the number n of quadrature points is doubled, i.e., if the step size h is halved, then the error for the Gaussian quadrature of orders one and two is reduced roughly by the factor 1/4 and 1/16, respectively.

TABLE 9.3. Gaussian quadrature for Example 9.6

m	$n = 1$	$n = 2$
1	0.02648051	0.00083949
2	0.00743289	0.00007054
4	0.00192729	0.00000489
8	0.00048663	0.00000031
16	0.00012197	0.00000002
32	0.00003051	0.00000000

9.4 Quadrature of Periodic Functions

We proceed by deriving the Euler–Maclaurin expansion.

Definition 9.23 *The* Bernoulli polynomials B_n *of degree n are defined recursively by* $B_0(x) := 1$ *and*

$$B_n' := B_{n-1}, \quad n \in \mathbb{N}, \tag{9.21}$$

with the normalization condition

$$\int_0^1 B_n(x)\,dx = 0, \quad n \in \mathbb{N}. \tag{9.22}$$

The rational numbers

$$b_n := n!\, B_n(0), \quad n = 0, 1, \ldots,$$

are called Bernoulli numbers.

The first Bernoulli polynomials are given by

$$B_0(x) = 1, \quad B_1(x) = x - \frac{1}{2}, \quad B_2(x) = \frac{1}{2}x^2 - \frac{1}{2}x + \frac{1}{12}.$$

We note that the normalization (9.22) is equivalent to

$$B_n(0) = B_n(1), \quad n = 2, 3, \ldots. \tag{9.23}$$

Lemma 9.24 *The Bernoulli polynomials have the symmetry property*

$$B_n(x) = (-1)^n B_n(1 - x), \quad x \in \mathbb{R}, \quad n = 0, 1, \ldots. \tag{9.24}$$

Proof. Obviously (9.24) holds for $n = 0$. Assume that (9.24) has been proven for some $n \geq 0$. Then, integrating (9.24), we obtain

$$B_{n+1}(x) = (-1)^{n+1} B_{n+1}(1 - x) + \beta_{n+1}$$

for some constant β_{n+1}. The condition (9.22) implies that $\beta_{n+1} = 0$, and therefore (9.24) is also valid for $n + 1$. □

Lemma 9.25 *The Bernoulli polynomials B_{2m+1}, $m = 1, 2, \ldots$, of odd degree have exactly three zeros in $[0, 1]$, and these zeros are at the points 0, $1/2$, and 1. The Bernoulli polynomials B_{2m}, $m = 0, 1, \ldots$, of even degree satisfy $B_{2m}(0) \neq 0$.*

Proof. From (9.23) and (9.24) we conclude that B_{2m+1} vanishes at the points 0, 1/2, and 1. We prove by induction that these are the only zeros of B_{2m+1} in $[0, 1]$. This is true for $m = 1$, since B_3 is a polynomial of degree three. Assume that we have proven that B_{2m+1} has only the three zeros 0, 1/2, and 1 in $[0, 1]$, and assume that B_{2m+3} has an additional zero α in $[0, 1]$. Because of the symmetry (9.24) we may assume that $\alpha \in (0, 1/2)$. Then, by Rolle's theorem, we conclude that B_{2m+2} has at least one zero in $(0, \alpha)$ and also at least one zero in $(\alpha, 1/2)$. Again by Rolle's theorem this implies that B_{2m+1} has a zero in $(0, 1/2)$, which contradicts the induction assumption.

From the zeros of B_{2m+1}, by Rolle's theorem it follows that B_{2m} has a zero in $(0, 1/2)$. Assume that $B_{2m}(0) = 0$. Then, by Rolle's theorem, B_{2m-1} has a zero in $(0, 1/2)$, which contradicts the first part of the lemma. □

By $\tilde{B}_n : \mathbb{R} \to \mathbb{R}$ we denote the periodic extension of the Bernoulli polynomial B_n; i.e., $\tilde{B}_n$ has period 1 and $\tilde{B}_n(x) = B_n(x)$ for $0 \leq x \leq 1$. The Fourier series of the periodic functions $\tilde{B}_n$ are given by

$$\tilde{B}_{2m}(x) = 2(-1)^{m-1} \sum_{k=1}^{\infty} \frac{\cos 2\pi kx}{(2\pi k)^{2m}} \tag{9.25}$$

and

$$\tilde{B}_{2m-1}(x) = 2(-1)^m \sum_{k=1}^{\infty} \frac{\sin 2\pi kx}{(2\pi k)^{2m-1}} \tag{9.26}$$

for $m = 1, 2, \ldots$. This follows from (9.21) and (9.22) and the elementary Fourier expansion for the piecewise linear function $\tilde{B}_1$ (see Problem 9.13).

Let $x_k = a + kh$, $k = 0, \ldots, n$, be an equidistant subdivision of the interval $[a, b]$ with step size $h = (b - a)/n$ and recall the definition of the trapezoidal sum

$$T_h(f) := h\left[\frac{1}{2}\, f(x_0) + f(x_1) + \cdots + f(x_{n-1}) + \frac{1}{2}\, f(x_n)\right]$$

for $f \in C[a, b]$.

Theorem 9.26 *Let $f : [a, b] \to \mathbb{R}$ be m times continuously differentiable for $m \geq 2$. Then we have the* Euler–Maclaurin expansion

$$\begin{aligned}\int_a^b f(x)\,dx = T_h(f) &- \sum_{j=1}^{[\frac{m}{2}]} \frac{b_{2j}h^{2j}}{(2j)!}\,[f^{(2j-1)}(b) - f^{(2j-1)}(a)] \\ &+(-1)^m h^m \int_a^b \tilde{B}_m\left(\frac{x-a}{h}\right) f^{(m)}(x)\,dx,\end{aligned} \tag{9.27}$$

where $\left[\frac{m}{2}\right]$ denotes the largest integer smaller than or equal to $\frac{m}{2}$.

Proof. Let $g \in C^m[0, 1]$. Then, by $m - 1$ partial integrations and using (9.23) we find that

$$\begin{aligned}\int_0^1 B_1(z)g'(z)\,dz &= \sum_{j=2}^{m}(-1)^j B_j(0)[g^{(j-1)}(1) - g^{(j-1)}(0)] \\ &\qquad -(-1)^m \int_0^1 B_m(z)g^{(m)}(z)\,dz.\end{aligned}$$

Combining this with the partial integration

$$\int_0^1 B_1(z)g'(z)\,dz = \frac{1}{2}\,[g(1) + g(0)] - \int_0^1 g(z)\,dz$$

and observing that the odd Bernoulli numbers vanish leads to

$$\begin{aligned}\int_0^1 g(z)\,dz = \frac{1}{2}\,[g(0) + g(1)] &- \sum_{j=1}^{[\frac{m}{2}]} \frac{b_{2j}}{(2j)!}\,[g^{(2j-1)}(1) - g^{(2j-1)}(0)] \\ &+(-1)^m \int_0^1 B_m(z)g^{(m)}(z)\,dz.\end{aligned}$$

Now we substitute $x = x_k + hz$ and $g(z) = f(x_k + hz)$ to obtain

$$\int_{x_k}^{x_{k+1}} f(x)\,dx = \frac{h}{2}\,[f(x_k) + f(x_{k+1})]$$

$$-\sum_{j=1}^{[\frac{m}{2}]} \frac{b_{2j}h^{2j}}{(2j)!}\,[f^{(2j-1)}(x_{k+1}) - f^{(2j-1)}(x_k)]$$

$$+(-1)^m h^m \int_{x_k}^{x_{k+1}} B_m\left(\frac{x-a}{h}\right) f^{(m)}(x)\,dx.$$

Finally, we sum the last equation for $k = 0, \dots, n-1$ to arrive at the Euler–Maclaurin expansion (9.27). □

For 2π-periodic continuous functions $f : \mathbb{R} \to \mathbb{R}$ the trapezoidal rule coincides with the *rectangular rule*

$$\int_0^{2\pi} f(x)\,dx \approx \frac{2\pi}{n} \sum_{k=1}^{n} f\left(\frac{2\pi k}{n}\right).$$

For its error

$$E_n(f) := \int_0^{2\pi} f(x)\,dx - \frac{2\pi}{n} \sum_{k=1}^{n} f\left(\frac{2\pi k}{n}\right)$$

we have the following corollary of the Euler–Maclaurin expansion.

Corollary 9.27 *Let $f : \mathbb{R} \to \mathbb{R}$ be $(2m+1)$-times continuously differentiable and 2π-periodic for $m \in \mathbb{N}$ and let $n \in \mathbb{N}$. Then for the error of the rectangular rule we have*

$$|E_n(f)| \le \frac{C}{n^{2m+1}} \int_0^{2\pi} |f^{(2m+1)}(x)|\,dx,$$

where

$$C := 2\sum_{k=1}^{\infty} \frac{1}{k^{2m+1}}.$$

Proof. From Theorem 9.26 we have that

$$E_n(f) = -\left(\frac{2\pi}{n}\right)^{2m+1} \int_0^{2\pi} \tilde{B}_{2m+1}\left(\frac{2\pi x}{n}\right) f^{(2m+1)}(x)\,dx,$$

and the estimate follows from the inequality

$$|\tilde{B}_{2m+1}(x)| \le 2\sum_{k=1}^{\infty} \frac{1}{(2\pi k)^{2m+1}}, \quad x \in \mathbb{R},$$

which is a consequence of (9.26). □

Corollary 9.27 illustrates why for periodic functions the simple rectangular rule is superior to any other quadrature rule (see Problem 9.12). Note that the rectangular rule can also be obtained by integrating the trigonometric interpolation polynomials of Theorems 8.24 and 8.25.

In the following theorem we give an example of derivative-free error estimates for numerical quadrature rules in the spirit of Davis [15]. They have the advantage that they do not need the computation of higher derivatives for the evaluation of the estimates. However, they require the integrand to be analytic, and their proofs need complex analysis.

Theorem 9.28 *Let $f : \mathbb{R} \to \mathbb{R}$ be analytic and 2π-periodic. Then there exists a strip $D = \mathbb{R} \times (-a, a) \subset \mathbb{C}$ with $a > 0$ such that f can be extended to a holomorphic and 2π-periodic bounded function $f : D \to \mathbb{C}$. The error for the rectangular rule can be estimated by*

$$|E_n(f)| \le \frac{4\pi M}{e^{na} - 1},$$

where M denotes a bound for the holomorphic function f on D.

Proof. Since $f : \mathbb{R} \to \mathbb{R}$ is analytic, at each point $x \in \mathbb{R}$ the Taylor expansion provides a holomorphic extension of f into some open disk in the complex plane with radius $r(x) > 0$ and center x. The extended function again has period 2π, since the coefficients of the Taylor series at x and at $x + 2\pi$ coincide for the 2π–periodic function $f : \mathbb{R} \to \mathbb{R}$. The disks corresponding to all points of the interval $[0, 2\pi]$ provide an open covering of $[0, 2\pi]$. Since $[0, 2\pi]$ is compact, a finite number of these disks suffices to cover $[0, 2\pi]$. Then we have an extension into a strip D with finite width $2a$ contained in the union of the finite number of disks. Without loss of generality we may assume that f is bounded on D.

From the residue theorem we have that

$$\int_{i\alpha}^{i\alpha+2\pi} \cot\frac{nz}{2}\, f(z)\, dz - \int_{-i\alpha}^{-i\alpha+2\pi} \cot\frac{nz}{2}\, f(z)\, dz = -\frac{4\pi i}{n} \sum_{k=1}^{n} f\left(\frac{2\pi k}{n}\right)$$

for each $0 < \alpha < a$. This implies that

$$\operatorname{Re} \int_{i\alpha}^{i\alpha+2\pi} i \cot\frac{nz}{2}\, f(z)\, dz = \frac{2\pi}{n} \sum_{k=1}^{n} f\left(\frac{2\pi k}{n}\right),$$

since by the Schwarz reflection principle, f enjoys the symmetry property $f(\bar{z}) = \overline{f(z)}$. By Cauchy's integral theorem we have

$$\operatorname{Re} \int_{i\alpha}^{i\alpha+2\pi} f(z)\, dz = \int_0^{2\pi} f(x)\, dx,$$

and combining the last two equations yields

$$E_n(f) = \operatorname{Re} \int_{i\alpha}^{i\alpha+2\pi} \left(1 - i\cot\frac{nz}{2}\right) f(z)\, dz$$

for all $0 < \alpha < a$. Now the estimate follows from

$$\left|1 - i \cot \frac{nz}{2}\right| \leq \frac{2}{e^{n\alpha} - 1}$$

for $\operatorname{Im} z = \alpha$ and then passing to the limit $\alpha \to a$. □

The estimate shows that for periodic analytic functions the rectangular rule is of exponential order; i.e., doubling the number of quadrature points doubles the number of correct digits in the approximate value for the integral.

9.5 Romberg Integration

We now proceed with describing the *extrapolation method* due to Richardson (1927). Its basic idea is to derive high-order approximation methods from simple low-order methods. It can be applied to a variety of formulae in numerical analysis, and its application to the Euler–Maclaurin expansion was suggested by Romberg in 1955.

Recall the composite trapezoidal rule

$$T_h^1(f) := h\left[\frac{1}{2}\, f(a) + \sum_{k=1}^{n-1} f(a+kh) + \frac{1}{2}\, f(b)\right]$$

with step size $h = (b-a)/n$. If f is four-times continuously differentiable, by the Euler–Maclaurin expansion from Theorem 9.26 we have an error representation of the form

$$\int_a^b f(x)\, dx = T_h^1(f) + \gamma_1 h^2 + O(h^4)$$

for some constant γ_1 depending on f but not on h. Hence, for half the step size, we have that

$$\int_a^b f(x)\, dx = T_{\frac{h}{2}}^1(f) + \gamma_1\, \frac{h^2}{4} + O(h^4).$$

From these two equations we can eliminate the terms containing h^2; i.e., we multiply the first equation by $-1/3$ and the second equation by $4/3$ and add both equations to obtain

$$\int_a^b f(x)\, dx = \frac{1}{3}\left[4T_{\frac{h}{2}}^1(f) - T_h^1(f)\right] + O(h^4).$$

Hence, the linear combination

$$T_h^2(f) := \frac{1}{3}\left[4T_{\frac{h}{2}}^1(f) - T_h^1(f)\right]$$

of the composite trapezoidal rule with step sizes h and $h/2$ leads to a quadrature formula with the improved error order $O(h^4)$. The quadrature $T_h^2(f)$ coincides with the composite Simpson's rule for the step size $h/2$.

If f is six-times continuously differentiable, by linearly combining the Euler–Maclaurin formulae for the step sizes h and $h/2$ we obtain an error representation of the form

$$\int_a^b f(x)\,dx = T_h^2(f) + \gamma_2 h^4 + O(h^6)$$

for some constant γ_2 depending only on f. From this and the corresponding formula

$$\int_a^b f(x)\,dx = T_{\frac{h}{2}}^2(f) + \gamma_2\,\frac{h^4}{16} + O(h^6)$$

for step size $h/2$, by eliminating the terms containing h^4 we obtain the quadrature formula

$$T_h^3(f) := \frac{1}{15}\left[16\,T_{\frac{h}{2}}^2(f) - T_h^2(f)\right]$$

with an error of order $O(h^6)$. Note that the actual numerical evaluation of $T_h^3(f)$ requires the values for the composite trapezoidal rule for the step sizes h, $h/2$, and $h/4$.

Obviously, this procedure can be repeated, and this leads to the sequence of *Romberg quadrature formulae*. Let

$$T_k^1(f) := T_{h_k}^1(f), \quad k = 0, 1, 2, \ldots,$$

be the trapezoidal sums for the step sizes $h_k := h/2^k$. Then for $m = 1, 2, \ldots$ the Romberg quadratures are recursively defined by

$$T_k^{m+1}(f) := \frac{1}{4^m - 1}\left[4^m T_{k+1}^m(f) - T_k^m(f)\right], \quad k = 0, 1, \ldots. \tag{9.28}$$

For the error we have the following theorem.

Theorem 9.29 *Let $f : [a,b] \to \mathbb{R}$ be $2m$-times continuously differentiable. Then for the Romberg quadratures we have the error estimate*

$$\left|\int_a^b f(x)\,dx - T_k^m(f)\right| \le C_m\,\|f^{(2m)}\|_\infty \left(\frac{h}{2^k}\right)^{2m}, \quad k = 0, 1, \ldots,$$

for some constant C_m depending on m.

Proof. By induction, we show that there exist constants $\gamma_{j,i}$ such that

$$\begin{aligned} &\left|\int_a^b f(x)\,dx - T_k^i(f) - \sum_{j=i}^{m-1} \gamma_{j,i}[f^{(2j-1)}(b) - f^{(2j-1)}(a)]\left(\frac{h}{2^k}\right)^{2j}\right| \\ &\quad \le \gamma_{m,i}\|f^{(2m)}\|_\infty \left(\frac{h}{2^k}\right)^{2m} \end{aligned} \tag{9.29}$$

for $i = 1, \dots, m$ and $k = 0, 1, \dots$. Here the sum on the left-hand side is set equal to zero for $i = m$. By the Euler–Maclaurin expansion this is true for $i = 1$ with $\gamma_{j,1} = b_{2j}/(2j)!$ for $j = 1, \dots, m-1$ and

$$\gamma_{m,1} = (b-a) \sup_{x \in [0,1]} |B_{2m}(x)|.$$

As an abbreviation we set

$$F_j := f^{(2j-1)}(b) - f^{(2j-1)}(a), \quad j = 1, \dots, m-1.$$

Assume that (9.29) has been shown for some $1 \le i < m$. Then, using (9.28), we obtain

$$\frac{4^i}{4^i - 1} \left[\int_a^b f(x)\,dx - T_{k+1}^i(f) - \sum_{j=i}^{m-1} \left(\frac{h}{2^{k+1}}\right)^{2j} \gamma_{j,i} F_j \right]$$

$$- \frac{1}{4^i - 1} \left[\int_a^b f(x)\,dx - T_k^i(f) - \sum_{j=i}^{m-1} \left(\frac{h}{2^k}\right)^{2j} \gamma_{j,i} F_j \right]$$

$$= \int_a^b f(x)\,dx - T_k^{i+1}(f) - \sum_{j=i+1}^{m-1} \left(\frac{h}{2^k}\right)^{2j} \gamma_{j,i+1} F_j,$$

where

$$\gamma_{j,i+1} = \frac{4^{i-j} - 1}{4^i - 1} \gamma_{j,i}, \quad j = i+1, \dots, m-1.$$

Now with the aid of the induction assumption we can estimate

$$\left| \int_a^b f(x)\,dx - T_k^{i+1}(f) - \sum_{j=i+1}^{m-1} \gamma_{j,i+1} \left(\frac{h}{2^k}\right)^{2j} \gamma_{j,i+1} F_j \right|$$

$$\le \gamma_{m,i+1} \|f^{(2m)}\|_\infty \left(\frac{h}{2^k}\right)^{2m},$$

where

$$\gamma_{m,i+1} = \frac{4^{i-m} + 1}{4^i - 1},$$

and the proof is complete. □

From Theorem 9.29 we conclude that the Romberg quadrature T_k^m integrates polynomials of degree less than or equal to $2m-1$ exactly. For $h = b - a$ the Romberg quadrature T_0^m uses $2^{m-1} + 1$ equidistant integration points. Therefore, T_0^1 coincides with the trapezoidal rule, T_0^2 with Simpson's rule, and T_0^3 with Milne's rule. Similarly, T_k^1, T_k^2, and T_k^3 correspond

to the composite trapezoidal rule, the composite Simpson's rule, and the composite Milne rule, respectively. For $m \geq 4$ the number of the quadrature points in T_0^m is greater than the degree of exactness. The Romberg formula T_0^4 uses nine quadrature points, and this is the number of quadrature points where the Newton–Cotes formulae start having negative weights.

Theorem 9.30 *The quadrature weights of the Romberg formulae are positive.*

Proof. We define recursively $Q_k^1 := 4T_{k+1}^1 - 2T_k^1$ and

$$Q_k^{m+1} := \frac{1}{4^m - 1}\left[2^{2m+1}T_{k+1}^m + 2T_k^m + 4^{m+1}Q_{k+1}^m\right] \tag{9.30}$$

for $k = 1, 2, \dots$ and $m = 1, 2 \dots$ and show by induction that

$$T_k^{m+1} = \frac{1}{4^m - 1}\left[T_k^m + Q_k^m\right]. \tag{9.31}$$

By the definition of Q_k^1 this is true for $m = 1$. We assume that (9.31) has been proven for some $m \geq 1$. Then, using the recursive definitions of T_k^m and Q_k^m and the induction assumption, we derive

$$\begin{aligned} T_k^{m+1} + Q_k^{m+1} &= \frac{4^{m+1}}{4^m - 1}\left[T_{k+1}^m + Q_{k+1}^m\right] - \frac{1}{4^m - 1}\left[4^m T_{k+1}^m - T_k^m\right] \\ &= 4^{m+1}T_{k+1}^{m+1} - T_k^{m+1} = (4^{m+1} - 1)T_k^{m+2}; \end{aligned}$$

i.e., (9.31) also holds for $m + 1$. Now, from (9.30) and (9.31), by induction with respect to m, it can be deduced that the weights of T_k^m are positive and that the weights of Q_k^m are nonnegative. □

Corollary 9.31 *For the Romberg quadratures we have convergence:*

$$\lim_{m\to\infty} T_k^m(f) = \int_a^b f(x)\,dx \quad \textit{and} \quad \lim_{k\to\infty} T_k^m(f) = \int_a^b f(x)\,dx$$

for all continuous functions f.

Proof. This follows from Theorems 9.29 and 9.30 and Corollary 9.11. □

For continuous functions, the trapezoidal sums converge as the step size tends to zero. This motivates us to consider a polynomial in h^2 interpolating the values $T_k^1(f), \dots, T_{k+m}^1(f)$ at the interpolation points $h_k^2, \dots, h_{k+m}^2$ and evaluate it at $h = 0$.

Theorem 9.32 *Denote by L_k^m the uniquely determined polynomial in h^2 of degree less than or equal to m with the interpolation property*

$$L_k^m(h_j^2) = T_j^1(f), \quad j = k, \dots, k + m.$$

Then the Romberg quadratures satisfy

$$T_k^{m+1}(f) = L_k^m(0). \tag{9.32}$$

Proof. Obviously, (9.32) is true for $m = 0$. Assume that it has been proven for $m - 1$. Then, using the Neville scheme from Theorem 8.9, we obtain

$$\begin{aligned} L_k^m(0) &= \frac{1}{h_{k+m}^2 - h_k^2} \left[-h_k^2 L_{k+1}^{m-1}(0) + h_{k+m}^2 L_k^{m-1}(0)\right] \\ &= \frac{1}{h_{k+m}^2 - h_k^2} \left[-h_k^2 T_{k+1}^m + h_{k+m}^2 T_k^m\right] \\ &= \frac{1}{4^m - 1} \left[4^m T_{k+1}^m - T_k^m\right] = T_k^{m+1}, \end{aligned}$$

establishing (9.32) for m. □

This interpretation of the Romberg quadrature as an *extrapolation method* in the sense of Richardson opens up the possibility of modifications using other than equidistant step sizes.

Table 9.4 gives the error between the exact value of the integral from Example 9.6 and its numerical approximation by the Romberg quadrature, exhibiting its fast convergence according to the error estimates of Theorem 9.29. Clearly, the first two columns of Table 9.4 have to coincide with Table 9.2.

TABLE 9.4. Romberg quadratures for Example 9.6

k	T_k^1	T_k^2	T_k^3	T_k^4
1	−0.05685282			
2	−0.01518615	−0.00129726		
4	−0.00387663	−0.00010679	−0.00002742	
8	−0.00097467	−0.00000735	−0.00000072	−0.00000030
16	−0.00024402	−0.00000047	−0.00000001	−0.00000000
32	−0.00006103	−0.00000003	−0.00000000	−0.00000000

We finish this section with the corresponding Table 9.5 for the integral

$$\int_0^1 \sqrt{x}\, dx = \frac{2}{3} \tag{9.33}$$

of a function that is not differentiable in all of the integration interval. Not surprisingly, the convergence is notably slower.

TABLE 9.5. Romberg quadratures for the integral (9.33)

k	T_k^1	T_k^2	T_k^3	T_k^4	T_k^5
1	0.166667				
2	0.063113	0.028595			
4	0.023384	0.010140	0.008910		
8	0.008536	0.003587	0.003151	0.003059	
16	0.003085	0.001268	0.001114	0.001082	0.001074
32	0.001108	0.000448	0.000394	0.000382	0.000380

9.6 Improper Integrals

We conclude this chapter with an example for the numerical integration of improper integrals and describe a class of quadrature rules for the integral

$$\int_0^1 f(x)\,dx$$

where the integrand f is sufficiently smooth in $(0, 1)$ but is allowed to have singularities at the endpoints $x = 0$ and $x = 1$ such that f is nonetheless integrable.

Let the function $w : [0, 2\pi] \to [0, 1]$ be bijective, strictly monotonically increasing, and infinitely differentiable. Then we can substitute $x = w(t)$ and consequently obtain

$$\int_0^1 f(x)\,dx = \int_0^{2\pi} g(t)\,dt,$$

where

$$g(t) := w'(t)\, f(w(t)), \quad 0 < t < 2\pi.$$

Now assume that the function w has derivatives

$$w^{(j)}(0) = w^{(j)}(2\pi) = 0, \quad j = 1, \dots, p-1, \tag{9.34}$$

and

$$w^{(p)}(0) \neq 0, \quad w^{(p)}(2\pi) \neq 0 \tag{9.35}$$

for some $p \in \mathbb{N}$. Then we may expect that the function g and some of its derivatives up to a certain order vanish at $t = 0$ and $t = 2\pi$; i.e., g can be considered as a sufficiently smooth 2π-periodic function, and the rectangular rule may be applied to the transformed integral. This yields the quadrature formula

$$\int_0^1 f(x)\,dx \approx \sum_{k=1}^{n-1} a_k f(x_k) \tag{9.36}$$

with the quadrature points and weights given by

$$x_k = w\left(\frac{2\pi k}{n}\right), \quad a_k = \frac{2\pi}{n}\, w'\left(\frac{2\pi k}{n}\right), \quad k = 1, \dots, n-1.$$

In addition, it is natural to require the symmetry property

$$w(t) = 1 - w(2\pi - t), \quad t \in [0, 2\pi]. \tag{9.37}$$

Then the quadrature points and weights have the symmetry

$$x_{n-k} = 1 - x_k, \quad a_{n-k} = a_k, \quad k = 1, \dots, n-1,$$

and from the assumptions (9.34) and (9.35), by Taylor's formula, it follows that they satisfy the inequalities

$$c_0 \left(\frac{k}{n}\right)^p \le x_k,\ 1 - x_{n-k} \le c_1 \left(\frac{k}{n}\right)^p, \quad k = 1, \dots, \left[\frac{n}{2}\right], \tag{9.38}$$

and

$$c_0 \left(\frac{k}{n}\right)^{p-1} \le a_k,\ a_{n-k} \le c_1 \left(\frac{k}{n}\right)^{p-1}, \quad k = 1, \dots, \left[\frac{n}{2}\right], \tag{9.39}$$

for some constants $0 < c_0 < c_1$ depending on the function w. From (9.38) it is obvious that the quadrature points are graded towards the two endpoints $x = 0$ and $x = 1$ of the integration interval.

For substitutions with the properties (9.34), (9.35), and (9.37), from the Euler–Maclaurin expansion applied to the integral over g we now will derive an estimate for the remainder term

$$E_n(f) := \int_0^1 f(x)\, dx - \sum_{k=1}^{n-1} a_k f(x_k).$$

For $q \in \mathbb{N}$ and $0 < \alpha \le 1$ by $S^{q,\alpha}$ we denote the linear space of q-times continuously differentiable functions $f : (0,1) \to \mathbb{R}$ for which

$$\sup_{0<x<1} [x(1-x)]^{j+1-\alpha}\, |f^{(j)}(x)| < \infty$$

for $j = 0, \dots, q$. On $S^{q,\alpha}$ we define the norm

$$\|f\|_{q,\alpha} := \max_{j=0,\dots,q}\ \sup_{0<x<1} [x(1-x)]^{j+1-\alpha}\, |f^{(j)}(x)|.$$

Then, clearly

$$|f^{(j)}(x)| \le \|f\|_{q,\alpha} [x(1-x)]^{\alpha-j-1}, \quad 0 < x < 1, \tag{9.40}$$

for $j = 0, \dots, q$.

Theorem 9.33 *Let $p \in \mathbb{N}$ and assume that w satisfies (9.34), (9.35), and (9.37). Further, let $q \in \mathbb{N}$ and $f \in S^{2q+1,\alpha}$ with $0 < \alpha \le 1$ such that*

$$2q + 1 < \alpha p \quad \textit{and} \quad 2q + 2 \le p.$$

Then the error in the quadrature formula (9.36) can be estimated by

$$|E_n(f)| \le \frac{C}{n^{2q+1}} \, \|f\|_{2q+1,\alpha}$$

with some constant C depending on w, α, and q.

Proof. For the derivatives of g we can write

$$g^{(r)}(t) = \sum_{j=0}^{r} u_j^r(t)\, f^{(j)}(w(t)), \quad r = 0, \ldots, 2q+1.$$

Then from

$$g^{(r+1)}(t) = \sum_{j=0}^{r} \left[u_j^r(t)\, w'(t)\, f^{(j+1)}(w(t)) + \frac{du_j^r(t)}{dt}\, f^{(j)}(w(t)) \right]$$

we derive the recursion formulae

$$u_j^{r+1}(t) = \begin{cases} \dfrac{du_0^r(t)}{dt}, & j = 0, \\[2ex] u_{j-1}^r(t)\, w'(t) + \dfrac{du_j^r(t)}{dt}, & j = 1, \ldots, r, \\[2ex] u_r^r(t)\, w'(t), & j = r+1, \end{cases} \tag{9.41}$$

for the coefficients u_j^r. In particular, we have

$$u_0^r(t) = w^{(r+1)}(t) \quad \text{and} \quad u_r^r(t) = [w'(t)]^{r+1}. \tag{9.42}$$

The functions u_j^r satisfy

$$u_j^r(t) = O([t(2\pi - t)])^{z_j^r}, \quad t(2\pi - t) \to 0, \tag{9.43}$$

for $r = 0, \ldots, 2q+1$ and $j = 0, \ldots, r$, where

$$z_j^r = p - 1 + jp - r.$$

For $j = 0$ and $j = r$ this is obvious from the assumption on w and (9.42), and for $j = 1, \ldots, r-1$ it follows by induction from the recursion formulae (9.41). Note that $z_j^r \ge 0$ because of the assumption $p \ge 2q + 2$.

Using (9.40) and the assumptions on w, we can estimate

$$|f^{(j)}(w(t))| \le C_1 \|f\|_{2q+1,\alpha} [t(2\pi - t)]^{(\alpha - j - 1)p}$$

for some constant C_1, and with the aid of (9.43), we further obtain that

$$|u_j^r(t)\, f^{(j)}(w(t))| \le C_2 \|f\|_{2q+1,\alpha}[t(2\pi - t)]^{\alpha p - r - 1}, \quad 0 < t < 2\pi, \tag{9.44}$$

for some constant C_2 and $r = 0, \dots, 2q+1$ and $j = 0, \dots, r$. From this, since $\alpha p > 2q+1$, we observe that for $r = 0, \dots, 2q$ the derivatives $g^{(r)}$ can be continuously extended from $(0, 2\pi)$ onto $[0, 2\pi]$ with values

$$g^{(r)}(0) = g^{(r)}(2\pi) = 0, \quad r = 0, \dots, 2q.$$

Furthermore, from (9.44) and the assumption $\alpha p > 2q+1$ we see that the integral of $g^{(2q+1)}$ over $[0, 2\pi]$ exists as an improper integral and

$$\int_0^{2\pi} |g^{(2q+1)}(t)|\, dt \le C_3 \|f\|_{2q+1,\alpha}$$

with some constant C_3 depending on w, α, and q. Now the statement follows from Corollary 9.27 of the Euler–Maclaurin expansion. Note that for the Euler–Maclaurin expansion (9.27) to be valid it obviously suffices that the integral of the error term exists as an improper integral. □

We proceed by describing a few examples for substitutions w (see Problem 9.19). In 1963 Korobov suggested the polynomial transformation

$$w_p(t) := \left[\int_0^{2\pi} [s(2\pi - s)]^{p-1} ds\right]^{-1} \int_0^t [s(2\pi - s)]^{p-1} ds\,. \tag{9.45}$$

The trigonometric transformation

$$w_p(t) := \left[\int_0^{2\pi} \sin^{p-1}\frac{s}{2}\, ds\right]^{-1} \int_0^t \sin^{p-1}\frac{s}{2}\, ds \tag{9.46}$$

with the special cases

$$w_1(t) = \frac{t}{2\pi}\,, \quad w_2(t) = \frac{1}{2}\left(1 - \cos\frac{t}{2}\right), \quad w_3(t) = \frac{1}{2\pi}\,(t - \sin t)$$

was proposed by Sidi [54]. Substitutions of the form

$$w_p(t) := \frac{t^p}{t^p + (2\pi - t)^p} \tag{9.47}$$

were considered in [40]. As a rule of thumb, these substitutions should not be used for p too large, say $p > 10$, because this may lead to overgrading and numerical difficulties with underflow. The substitutions

$$w(t) = \left[\int_0^{2\pi} \exp\left(-\frac{\pi}{s} - \frac{\pi}{2\pi - s}\right) ds\right]^{-1} \int_0^t \exp\left(-\frac{\pi}{s} - \frac{\pi}{2\pi - s}\right) ds$$

and

$$w(t) = \frac{\exp\left(-\frac{2\pi}{t}\right)}{\exp\left(-\frac{2\pi}{t}\right) + \exp\left(-\frac{2\pi}{2\pi - t}\right)}$$

with zeros of infinite order at the endpoints, which were suggested by Iri, Moriguti, and Takesawa [58] and by Sag and Szekeres [52], respectively, also suffer from this drawback.

As a numerical example we consider the improper integral

$$\int_0^1 \frac{1}{\sqrt{x}}\, dx = 2. \tag{9.48}$$

Table 9.6 gives the error between the exact value and the numerical approximation obtained by using the substitution (9.47).

TABLE 9.6. Numerical quadrature for the integral (9.48)

n	$p = 3$	$p = 4$	$p = 5$	$p = 6$
8	0.07012542	−0.06064201	−0.22007377	−0.42795942
16	0.02849925	0.00455233	−0.00438402	−0.01896018
32	0.00992273	0.00129852	0.00011279	−0.00003394
64	0.00347755	0.00032530	0.00002117	−0.00000019
128	0.00122386	0.00008137	0.00000382	−0.00000001

Problems

9.1 Show that the error for the composite trapezoidal rule can be expressed in the form

$$\int_a^b f(x)\,dx - T_h(f) = -\int_a^b K_T(x) f''(x)\,dx,$$

where the so-called *Peano kernel* K_T is given by

$$K_T(x) = \frac{1}{2}\,(x - x_{k-1})(x_k - x), \quad x_{k-1} \le x \le x_k,$$

for $k = 1, \dots, n$. Use this error representation for an alternative proof of Theorem 9.7.

9.2 Show that the error for the composite Simpson's rule can be expressed in the form

$$\int_a^b f(x)\,dx - S_h(f) = -\int_a^b K_S(x) f^{(4)}(x)\,dx,$$

where the Peano kernel K_S is given by

$$K_S(x) := \begin{cases} \frac{h}{18}(x - x_{k-2})^3 - \frac{1}{24}(x - x_{k-2})^4, & x_{k-2} \le x \le x_{k-1}, \\ \frac{h}{18}(x_k - x)^3 - \frac{1}{24}(x_k - x)^4, & x_{k-1} \le x \le x_k, \end{cases}$$

for $k = 2, 4, \dots, n$. Use this error representation for an alternative proof of Theorem 9.8.

9.3 For Newton's three-eights rule prove the error representation

$$\int_a^b f(x)\,dx - \frac{b-a}{8}\,[f(a) + 3f(a+h) + 3f(b-h) + f(b)] = -\frac{3h^5}{80}\,f^{(4)}(\xi)$$

with some ξ in $[a, b]$ and $h = (b-a)/3$.

9.4 Show that the weight a_4 for the Newton–Cotes formula of order eight is negative.

9.5 For the remainder E_n of the Newton–Cotes formula of order n on the interval $[-1, 1]$, applied to the Chebyshev polynomial T_{n+1}, show that

$$E_n(T_{n+1}) = \frac{(n+1)!\,4^{n+1}}{n^{n+2}} \int_0^n \binom{z}{n+1} dz, \quad n \in \mathbb{N}.$$

From this conclude that if n odd, then

$$\|E_n\|_\infty \ge |E_n(T_{n+1})| \ge \gamma_n,$$

where

$$\gamma_n = \frac{(n-1)!\,4^{n+1}}{3\,n^{n+2}} \to \infty, \quad n \to \infty.$$

Hint: Use Theorem 8.10 and show that

$$\int_0^n \binom{z}{n+1} dz = -2 \int_0^1 \binom{z}{n+2} dz$$

for n odd.

9.6 Compute the weights for the polynomial interpolatory quadratures with equidistant quadrature points

$$x_k = a + (k+1)\,\frac{b-a}{n+2}, \quad k = 0, 1, \dots, n,$$

for $n = 0, 1, 2$ and obtain representations of the quadrature errors. These formulae are called *open Newton–Cotes quadratures*, since the two endpoints a and b are omitted.

9.7 For $n \in \mathbb{N}$, a quadrature formula of the form

$$\int_a^b f(x)\,dx \approx \frac{b-a}{n}\sum_{k=1}^{n} f(x_k)$$

with distinct quadrature points $x_1, \ldots, x_n \in [a, b]$ and equal weights is called a *Chebyshev quadrature* if it integrates polynomials in P_n exactly. Find the Chebyshev quadratures for $n = 1, 2, 3, 4$. (Chebyshev quadratures exist only for $n < 8$.)

9.8 Show that there exists no polynomial interpolatory quadrature of order n that integrates polynomials of degree $2n + 2$ exactly.

9.9 The *Chebyshev polynomial* of the second kind U_n of degree n is defined by

$$U_n(x) := \frac{\sin((n+1)\arccos x)}{\sin(\arccos x)}, \quad -1 \le x \le 1.$$

Show that $U_0(x) = 1$, $U_1(x) = 2x$, and

$$U_{n+1}(x) + U_{n-1}(x) = 2xU_n(x), \quad n = 1, 2, \ldots.$$

Prove the orthogonality relation

$$\int_{-1}^{1} \sqrt{1-x^2}\, U_n(x) U_m(x)\,dx = \frac{\pi}{2}\,\delta_{nm}.$$

9.10 Show that the quadrature points and quadrature weights for the *Gauss–Chebyshev quadrature* of order $n - 1$ for the integral

$$\int_{-1}^{1} \sqrt{1-x^2}\, f(x)\,dx$$

are given by

$$x_k = \cos\frac{k+1}{n+1}\,\pi$$

and

$$a_k = \frac{\pi}{n+1}\,\sin^2\frac{k+1}{n+1}\,\pi$$

for $k = 0, \ldots, n-1$.

9.11 Find the quadrature weights a_0, a_1, a_2, a_3, and the (remaining) quadrature points x_1, x_2 of a quadrature formula of the form

$$\int_{-1}^{1} f(x)\,dx \approx a_0 f(-1) + a_1 f(x_1) + a_2 f(x_2) + a_3 f(1)$$

that is exact for all polynomials in P_5. (This is an example of a *Gauss–Lobatto quadrature*, i.e., a Gauss quadrature with two preassigned quadrature points.)

Find the quadrature weights a_0, a_1, a_2, and the (remaining) quadrature points x_1, x_2 of a quadrature formula of the form

$$\int_{-1}^{1} f(x)\,dx \approx a_0 f(-1) + a_1 f(x_1) + a_2 f(x_2)$$

that is exact for all polynomials in P_4. (This is an example of a *Gauss–Radau quadrature*, i.e., a Gauss quadrature with one preassigned quadrature point.)

9.12 By approximating the integral

$$\int_0^{2\pi} \frac{1}{5 - 4\cos x}\, dx = \frac{2\pi}{3}$$

by the rectangular rule and Simpson's rule convince yourself of the superiority of the rectangular rule for periodic functions.

9.13 Verify the Fourier series (9.25) and (9.26) for the periodic Bernoulli polynomials.

9.14 For the Bernoulli polynomials show that the series

$$\sum_{n=0}^{\infty} B_n(x)t^n = \frac{te^{xt}}{e^t - 1}$$

is absolutely and locally uniformly convergent for all $x \in [0, 1]$ and all $t \in (-1, 1)$.

9.15 Derive a quadrature formula by integrating the interpolating cubic spline from Theorem 8.30 and discuss its relation to the Euler–Maclaurin expansion.

9.16 Write a computer program for Romberg integration and test it for various examples.

9.17 Calculate the weights of the Romberg quadratures T_k^3 and T_k^4.

9.18 Show that the Richardson extrapolation for the midpoint rule (9.20) leads to nonnegative quadrature weights.

9.19 Show that the functions (9.45), (9.46), and (9.47) are strictly monotonically increasing, infinitely differentiable, and map $[0, 2\pi]$ onto $[0, 1]$ such that (9.34), (9.35), and (9.37) are satisfied.

9.20 Write a computer program for the numerical quadrature (9.36) using the substitution (9.47) and test it for various examples.

10

Initial Value Problems

Historically, the study of differential equations originated in the beginnings of calculus with Newton and Leibniz in the seventeenth century and is closely interwoven with the general development of mathematics. To a substantial degree, the central role of differential equations within mathematics is due to the fact that many important problems in science and engineering are modeled by differential equations.

This chapter will be devoted to an introduction to the basic numerical approximation methods for initial value problems for ordinary differential equations. For a more comprehensive study we refer to [13, 33, 42, 46, 55]. Analogous to the need for numerical quadrature formulas, numerical methods for the approximate solution of ordinary differential equations are necessary, since in general, no explicit solutions of the differential equation will be known, despite the fact that there exists a broad range of analytical solution methods for special classes of ordinary differential equations. In addition, the functions and data involved in the differential equation problem quite often will be available only at discrete points. However, we would like to emphasize that despite the availability of numerical methods the study of elementary analytical methods for the solution of ordinary differential equations remains worthwhile, since it provides a first step into gaining insight into the general structure of differential equations.

A solid foundation for numerical approximation methods for differential equations, including their convergence and error analysis, requires as a prerequisite results on the existence and uniqueness of the solution to the problem to be approximately solved. Therefore, in Section 10.1 we will begin with proving the fundamental Picard–Lindelöf existence and unique-

ness theorem for initial value problems. In Section 10.2 we will describe some variants of the simplest method for the numerical solution of initial value problems, which was first used by Euler. These methods are special cases of so-called single-step methods, for which we will give a convergence and error analysis in Section 10.3. This section also includes a short discussion of the Runge–Kutta method as the most widely used single-step method. The final section, Section 10.4, is concerned with the description and analysis of multistep methods.

We wish to note explicitly that this chapter is also meant to serve as an application of some of the material provided in Chapters 8 and 9 on interpolation and numerical integration.

10.1 The Picard–Lindelöf Theorem

Definition 10.1 *Let $G \subset \mathbb{R}^2$ be a domain and $f : G \to \mathbb{R}$. A continuously differentiable function $u : [a,b] \to \mathbb{R}$ is called a solution of the* ordinary differential equation of the first order

$$u' = f(x,u) \tag{10.1}$$

if $(x,u(x)) \in G$ and $u'(x) = f(x,u(x))$ for all $x \in [a,b]$.

Geometrically speaking, the differential equation (10.1) defines a field of directions on G. Solving the differential equation means looking for functions whose graphs match this field of directions.

Systems of ordinary differential equations can be included in the discussion as follows. If $G \subset \mathbb{R}^{n+1}$ is a domain and $f : G \to \mathbb{R}^n$, then a continuously differentiable function $u : [a,b] \to \mathbb{R}^n$ is called a solution of the system of ordinary differential equations of the first order

$$u' = f(x,u)$$

if $(x,u(x)) \in G$ and $u'(x) = f(x,u(x))$ for all $x \in [a,b]$. More explicitly, this system reads

$$\begin{aligned} u_1' &= f_1(x,u_1,\dots,u_n) \\ u_2' &= f_2(x,u_1,\dots,u_n) \\ &\cdot \\ &\cdot \\ &\cdot \\ u_n' &= f_n(x,u_1,\dots,u_n). \end{aligned}$$

for $u = (u_1,\dots,u_n)^T$ and $f = (f_1,\dots,f_n)^T$. Each ordinary differential equation

$$u^{(n)} = f(x,u,u',\dots,u^{(n-1)})$$

of order n is equivalent to the system

$$u_1' = u_2, \quad u_2' = u_3, \quad \dots, \quad u_{n-1}' = u_n, \quad u_n' = f_n(x, u_1, \dots, u_n)$$

via $u_1 = u$, $u_2 = u', \dots, u_n = u^{(n-1)}$. Therefore, in principle, considering only differential equations of the first order is no loss of generality.

From the wide field of applications we sketch only the following two simple examples.

Example 10.2 By Newton's law, the differential equation of the second order

$$mu'' = f(t, u)$$

describes the motion of an object of mass m subject to the external force $f(t, u)$ depending on the location u of the object and the time t. Given an initial location u_0 and an initial velocity u_0' at the initial time $t = 0$, one wants to find the position $u(t)$ of the object for all times $t \geq 0$. □

Example 10.3 Let $p = p(t)$ describe the population of a species of animals or plants at time t. If $r(t, p)$ denotes the growth rate given by the difference between the birth and death rate depending on the time t and the size p of the population, then an isolated population satisfies the differential equation

$$\frac{dp}{dt} = r(t, p).$$

The simplest model $r(t, p) = ap$, where a is a positive constant, leads to

$$\frac{dp}{dt} = ap$$

with the explicit solution $p(t) = p_0 e^{a(t-t_0)}$. Such an exponential growth is realistic only if the population is not too large. The modified model

$$\frac{dp}{dt} = ap - bp^2$$

with positive constants a and b contains a correction term that slows down the growth rate for large populations and is known as the *Verhulst equation*. It was introduced by Verhulst in 1938 as a model for the growth of the human population. In general, for a given growth rate r one wants to determine the development of the population $p(t)$ in time for a given initial population p_0 at time $t = t_0$. □

Both examples are typical initial value problems: Find a solution of a differential equation that attains a given initial value at a given initial time. This notion is made more precise by the following definition.

Definition 10.4 *The* initial value problem *for the ordinary differential equation*

$$u' = f(x, u) \tag{10.2}$$

consists in finding a continuously differentiable solution u satisfying the initial condition

$$u(x_0) = u_0 \tag{10.3}$$

for a given initial point x_0 and a given initial value u_0.

The existence and uniqueness of a solution to such an initial value problem are settled through the following fundamental theorem.

Theorem 10.5 (Picard–Lindelöf) *Let $G \in \mathbb{R}^{n+1}$ be a domain and let $f : G \to \mathbb{R}^n$ be a continuous function satisfying a* Lipschitz condition

$$\|f(x, u) - f(x, v)\| \le L\|u - v\| \tag{10.4}$$

for all (x, u), $(x, v) \in G$ and some constant $L > 0$, which is called the Lipschitz constant. *Then for each initial data pair $(x_0, u_0) \in G$ there exists an interval $[x_0 - a, x_0 + a]$ with $a > 0$ such that the initial value problem (10.2)–(10.3) has a unique solution in this interval.*

Proof. Firstly, we transform the initial value problem equivalently into the *Volterra integral equation*

$$u(x) = u_0 + \int_{x_0}^{x} f(\xi, u(\xi))\, d\xi. \tag{10.5}$$

Clearly, if u solves the initial value problem, then it follows by integrating the differential equation and using the initial condition that u also solves the integral equation. Conversely, if u is a continuous solution of the integral equation, then by differentiating the integral equation it follows that u is continuously differentiable and satisfies the differential equation. Inserting $x = x_0$ in (10.5) shows that the initial condition is fulfilled.

For solving the Volterra integral equation we now can employ Banach's fixed point Theorem 3.45. Since G is open, we can choose a bounded domain D such that $(x_0, u_0) \in D$ and $\overline{D} \subset G$. Denote by M a bound on the continuous function $f : \overline{D} \to \mathbb{R}^n$; i.e.,

$$\|f(x, u)\| \le M, \quad (x, u) \in \overline{D}.$$

Since D is open, we can choose $a > 0$ such that the closed rectangle

$$B := \{(x, u) \in \mathbb{R}^{n+1} : |x - x_0| \le a,\ \|u - u_0\| \le Ma\}$$

is contained in D. Consider the Banach space $C[x_0 - a, x_0 + a]$ of continuous functions $u : [x_0 - a, x_0 + a] \to \mathbb{R}^n$ furnished with the maximum norm

$$\|u\|_\infty := \max_{|x - x_0| \le a} \|u(x)\|$$

in terms of the chosen norm $\|\cdot\|$ on $\mathbb{R}^n$. Each solution u of the integral equation satisfies (see (6.1))

$$\|u(x) - u_0\| = \left\| \int_{x_0}^{x} f(\xi, u(\xi))\, d\xi \right\| \le Ma, \quad |x - x_0| \le a,$$

that is,

$$\|u - u_0\|_\infty \le Ma,$$

which implies that the solution remains within the rectangle B. We consider the closed subset

$$U := \{u \in C[x_0 - a, x_0 + a] : \|u - u_0\|_\infty \le Ma\}$$

of the Banach space $C[x_0 - a, x_0 + a]$ and note that by Remark 3.40 the set U is complete. On U we define an operator $A : U \to U$ by setting

$$(Au)(x) := u_0 + \int_{x_0}^{x} f(\xi, u(\xi))\, d\xi, \quad |x - x_0| \le a.$$

The operator A indeed maps U into itself, since the function Au is continuous and satisfies $\|Au - u_0\|_\infty \le Ma$. With the aid of the Lipschitz condition (10.4) and using (6.1) we can estimate

$$\begin{aligned}\|(Au)(x) - (Av)(x)\| &= \left\| \int_{x_0}^{x} [f(\xi, u(\xi)) - f(\xi, v(\xi))]\, d\xi \right\| \\ &\le L \int_{x_0}^{x} \|u(\xi) - v(\xi)\|\, d\xi \le La\|u - v\|_\infty\end{aligned}$$

for all $|x - x_0| \le a$. Hence

$$\|Au - Av\|_\infty \le La\|u - v\|_\infty$$

for all $u, v \in U$. Now we choose a such that $a < 1/L$. Then $A : U \to U$ is a contraction operator, and the Banach fixed point theorem ensures a unique fixed point of A, i.e., a unique solution of the integral equation (10.5) in the interval $[x_0 - a, x_0 + a]$. □

Exploiting the fact that in Theorem 10.5 the width a of the interval is determined by the Lipschitz constant L, which is independent of the initial point (x_0, u_0), one can assure *global* existence of the solution; i.e., the solution to the initial value problem exists and is unique until it leaves the domain G of definition for the differential equation.

Note that on a convex domain each function that is continuously differentiable with respect to u satisfies a Lipschitz condition (see the mean value Theorem 6.7).

Corollary 10.6 *Under the assumptions of Theorem 10.5, the sequence* (u_ν) *defined by* $u_0(x) = u_0$ *and*

$$u_{\nu+1}(x) := u_0 + \int_{x_0}^{x} f(\xi, u_\nu(\xi))\, d\xi, \quad |x - x_0| \le a, \quad \nu = 0, 1, \ldots, \tag{10.6}$$

converges as $\nu \to \infty$ *uniformly on* $[x_0 - a, x_0 + a]$ *to the unique solution* u *of the initial value problem. We have the a posteriori error estimate*

$$\|u - u_\nu\|_\infty \le \frac{La}{1 - La} \|u_\nu - u_{\nu-1}\|_\infty, \quad \nu = 1, 2, \ldots.$$

Proof. This follows from Theorem 3.46. □

Example 10.7 Consider the initial value problem

$$u' = x^2 + u^2, \quad u(0) = 0,$$

on $G = (-0.5, 0.5) \times (-0.5, 0.5)$. For $f(x, u) := x^2 + u^2$ we have

$$|f(x, u)| \le 0.5$$

on G. Hence for any $a < 0.5$ and $M = 0.5$ the rectangle B from the proof of Theorem 10.5 satisfies $B \subset G$. Furthermore, we can estimate

$$|f(x, u) - f(x, v)| = |u^2 - v^2| = |(u + v)(u - v)| \le |u - v|$$

for all (x, u), $(x, v) \in G$; i.e., f satisfies a Lipschitz condition with Lipschitz constant $L = 1$. Thus in this case the contraction number in the Picard–Lindelöf theorem is given by $La < 0.5$.

Here, the iteration (10.6) reads

$$u_{\nu+1}(x) = \int_0^x [\xi^2 + u_\nu^2(\xi)]\, d\xi.$$

Starting with $u_0(x) = 0$ we first compute

$$u_1(x) = \int_0^x \xi^2 d\xi = \frac{x^3}{3},$$

and from Corollary 10.6 we have the error estimate

$$\|u - u_1\|_\infty \le \|u_1 - u_0\|_\infty = \frac{1}{24} = 0.041\ldots.$$

The second iteration yields

$$u_2(x) = \int_0^x \left[\xi^2 + \frac{\xi^6}{9}\right] d\xi = \frac{x^3}{3} + \frac{x^7}{63}$$

with the error estimate

$$\|u - u_2\|_\infty \le \|u_2 - u_1\|_\infty = \frac{1}{63 \cdot 2^7} = 0.00012\ldots,$$

and the third iteration yields

$$u_3(x) = \int_0^x \left[\xi^2 + \frac{\xi^6}{9} + \frac{2\xi^{10}}{189} + \frac{\xi^{14}}{3969}\right] d\xi = \frac{x^3}{3} + \frac{x^7}{63} + \frac{2x^{11}}{2079} + \frac{x^{15}}{59535}$$

with the error estimate

$$\|u - u_3\|_\infty \le \|u_3 - u_2\|_\infty = \frac{1}{2079 \cdot 2^{10}} + \frac{1}{59535 \cdot 2^{15}} = 0.00000047\ldots.$$

In this example three steps of the Picard–Lindelöf iteration give eight decimal places of accuracy. However, the example is not typical, since in general, the integrations required in each iteration step will not be available explicitly as in the present case. □

10.2 Euler's Method

In the sequel we confine our presentation to the initial value problem for a differential equation of the first order. The generalization to systems and henceforth to equations of higher order is straightforward. We shall always tacitly assume that the assumptions of the Picard–Lindelöf Theorem 10.5 are satisfied.

The following simple method for the numerical solution of the initial value problem

$$u' = f(x, u), \quad u(x_0) = u_0, \tag{10.7}$$

was first used by Euler. Given a step size $h > 0$, it consists in replacing the derivative $u' = f(x, u)$ throughout the interval $[x_0, x_0 + h]$ by the derivative $u'_0 = f(x_0, u_0)$ at the initial point, i.e., geometrically speaking, by replacing the solution by its tangent line at the initial point x_0. This leads to the approximation

$$u_1 = u_0 + hf(x_0, u_0) \tag{10.8}$$

for the value $u(x_1)$ of the exact solution at the point $x_1 = x_0 + h$. Repeating this procedure leads to the Euler method as described in the following definition. For obvious reason, this method is also known as the *polygon method*, since it approximates the exact solution curve by a polygon.

Definition 10.8 *The* Euler method *for the numerical solution of the initial value problem (10.7) constructs approximations u_j to the exact solution $u(x_j)$ at the equidistant grid points*

$$x_j := x_0 + jh, \quad j = 1, 2, \ldots,$$

with step size h by

$$u_{j+1} := u_j + hf(x_j, u_j), \quad j = 0, 1, \dots .$$

Example 10.9 Consider the initial value problem

$$u' = x^2 + u^2, \quad u(0) = 0,$$

from Example 10.7. Table 10.1 gives the difference between the exact solution as computed by the Picard–Lindelöf iterations in Example 10.7 and the approximate solution obtained by Euler's method for various step sizes h. We observe a linear convergence as $h \to 0$. □

TABLE 10.1. Numerical example for the Euler method

x	$h = 0.1$	$h = 0.01$	$h = 0.001$	$h = 0.0001$
0.1	0.000333	0.000048	0.000005	0.000000
0.2	0.001667	0.000197	0.000020	0.000002
0.3	0.004003	0.000446	0.000045	0.000005
0.4	0.007357	0.000798	0.000080	0.000008
0.5	0.011769	0.001258	0.000127	0.000013

There are three different interpretations of the approximation formula of Euler's method:

1. Replace the derivative by the difference quotient

$$\frac{u(x_1) - u(x_0)}{h} \approx u'(x_0) = f(x_0, u_0)$$

and solve for $u(x_1)$.

2. Integrate in the equivalent integral equation (10.5), i.e., in

$$u(x_1) = u(x_0) + \int_{x_0}^{x_1} f(\xi, u(\xi))\, d\xi$$

approximately by the rectangular rule

$$\int_{x_0}^{x_1} f(\xi, u(\xi))\, d\xi \approx hf(x_0, u_0).$$

3. Use Taylor's formula

$$u(x_1) = u(x_0) + hu'(x_0) + \frac{h^2}{2} u''(x_0 + \theta h)$$

with $0 < \theta < 1$ and neglect the remainder term; i.e., approximate $u(x_1) \approx u(x_0) + hu'(x_0)$.

Each of these three interpretations opens up possibilities for improvements of Euler's method. For example, instead of the rectangular rule we can use the more accurate trapezoidal rule

$$\int_{x_0}^{x_1} f(\xi, u(\xi))\, d\xi \approx \frac{h}{2}\,[f(x_0, u(x_0)) + f(x_1, u(x_1))],$$

which yields

$$u_1 = u_0 + \frac{h}{2}\,[f(x_0, u_0) + f(x_1, u_1)]. \tag{10.9}$$

Repeating this procedure leads to the following method.

Definition 10.10 *The* implicit Euler method *for the numerical solution of the initial value problem (10.7) constructs approximations u_j to the exact solution $u(x_j)$ at the equidistant grid points*

$$x_j := x_0 + jh, \quad j = 1, 2, \dots,$$

with step size h by

$$u_{j+1} = u_j + \frac{h}{2}\,[f(x_j, u_j) + f(x_{j+1}, u_{j+1})], \quad j = 0, 1, \dots.$$

This method is called an *implicit method*, since determining u_{j+1} requires the solution of an equation that in general is nonlinear. In contrast, the Euler method of Definition 10.8 is an *explicit method*, since it provides an explicit expression for the computation of u_{j+1}.

Remark 10.11 *The nonlinear equations of the implicit Euler method can be solved by successive approximations, provided that the Lipschitz constant L for f and the step size h satisfy $Lh < 2$.*

Proof. We have to solve equation (10.9) for u_1. Setting

$$g(u) := u_0 + \frac{h}{2}\,[f(x_0, u_0) + f(x_1, u)]$$

we can rewrite (10.9) as the fixed point equation $u_1 = g(u_1)$. The function g is a contraction, since

$$|g(u) - g(v)| = \frac{h}{2}\,|f(x_1, u) - f(x_1, v)| \le \frac{hL}{2}\,|u - v|,$$

and therefore the assertion follows from Theorem 3.46. □

Since the solution of the nonlinear equation (10.9) will deliver only an approximation to the solution of the initial value problem, there is no need to solve (10.9) with high accuracy. Using the approximate value from the explicit Euler method as a starting point and carrying out only one iteration, we arrive at the following method.

Definition 10.12 *The* predictor corrector *method for the Euler method for the numerical solution of the initial value problem (10.7), also known as the* improved Euler method *or* Heun method, *constructs approximations* u_j *to the exact solution* $u(x_j)$ *at the equidistant grid points*

$$x_j := x_0 + jh, \quad j = 1, 2, \ldots,$$

by

$$u_{j+1} := u_j + \frac{h}{2}\,[f(x_j, u_j) + f(x_{j+1}, u_j + hf(x_j, u_j))], \quad j = 0, 1, \ldots.$$

Example 10.13 Consider again the initial value problem from Example 10.7. Table 10.2 gives the difference between the exact solution as computed by the Picard–Lindelöf iterations and the approximate solution obtained by the improved Euler method for various step sizes h. We observe quadratic convergence as $h \to 0$. □

TABLE 10.2. Numerical example for the improved Euler method

x	$h = 0.1$	$h = 0.01$	$h = 0.001$
0.1	−0.00016667	−0.00000167	−0.00000002
0.2	−0.00033326	−0.00000333	−0.00000003
0.3	−0.00049955	−0.00000500	−0.00000005
0.4	−0.00066530	−0.00000668	−0.00000007
0.5	−0.00083027	−0.00000837	−0.00000009

In the following section we will show that the Euler method and the improved Euler method are convergent with convergence order one and two, respectively, as observed in the special cases of Examples 10.9 and 10.13.

10.3 Single-Step Methods

We generalize the Euler methods into more general single-step methods by the following definition.

Definition 10.14 Single-step methods *for the approximate solution of the initial value problem*

$$u' = f(x, u), \quad u(x_0) = u_0,$$

construct approximations u_j *to the exact solution* $u(x_j)$ *at the equidistant grid points*

$$x_j := x_0 + jh, \quad j = 1, 2, \ldots,$$

with step size h by

$$u_{j+1} := u_j + h\varphi(x_j, u_j; h), \quad j = 0, 1, \ldots,$$

where the function $\varphi : G \times (0, \infty) \to \mathbb{R}$ *is given in terms of the right-hand side* $f : G \to \mathbb{R}$ *of the differential equation.*

Example 10.15 The Euler method and the improved Euler method are single-step methods with

$$\varphi(x, u; h) = f(x, u) \tag{10.10}$$

and

$$\varphi(x, u; h) = \frac{1}{2}\,[f(x, u) + f(x + h, u + hf(x, u))], \tag{10.11}$$

respectively. □

The function φ describes how the differential equation

$$u' = f(x, u)$$

is approximated by the difference equation

$$\frac{1}{h}\,[u(x + h) - u(x)] = \varphi(x, u; h).$$

From a reasonable approximation we expect that the exact solution to the initial value problem approximately satisfies the difference equation. Hence,

$$\frac{1}{h}\,[u(x + h) - u(x)] - \varphi(x, u; h) \to 0, \quad h \to 0,$$

must be fulfilled for the exact solution u. We also expect that the order of this convergence will influence the accuracy of the approximate solution. These considerations are made more precise by the following definition.

Definition 10.16 *For each* $(x, u) \in G$ *denote by* $\eta = \eta(\xi)$ *the unique solution to the initial value problem*

$$\eta' = f(\xi, \eta), \quad \eta(x) = u,$$

with initial data (x, u). *Then*

$$\Delta(x, u; h) := \frac{1}{h}\,[\eta(x + h) - \eta(x)] - \varphi(x, u; h)$$

is called the local discretization error. *The single-step method is called* consistent *(with the initial value problem) if*

$$\lim_{h \to 0} \Delta(x, u; h) = 0$$

uniformly for all $(x, u) \in G$, *and it is said to have* consistency order p *if*

$$|\Delta(x, u; h)| \leq Kh^p$$

for all $(x, u) \in G$, *all* $h > 0$, *and some constant* K.

Without loss of generality, in the sequel we always will assume that f (and later also derivatives of f) are uniformly continuous and bounded on G. This can always be achieved by reducing G to a smaller domain.

Theorem 10.17 *A single-step method is consistent if and only if*

$$\lim_{h \to 0} \varphi(x, u; h) = f(x, u)$$

uniformly for all $(x, u) \in G$.

Proof. Since we assume f to be bounded, we have

$$\eta(x+t) - \eta(x) = \int_0^t \eta'(x+s)\,ds = \int_0^t f(x+s, \eta(x+s))\,ds \to 0, \quad t \to 0,$$

uniformly for all $(x, u) \in G$. Therefore, since we also assume that f is uniformly continuous, it follows that

$$\frac{1}{h}\left|\int_0^h [\eta'(x+t) - \eta'(x)]\,dt\right| \le \max_{0 \le t \le h} |\eta'(x+t) - \eta'(x)|$$

$$= \max_{0 \le t \le h} |f(x+t, \eta(x+t)) - f(x, \eta(x))| \to 0, \quad h \to 0,$$

uniformly for all $(x, u) \in G$. From this we obtain that

$$\Delta(x, u; h) + \varphi(x, u; h) - f(x, u) = \frac{1}{h}\,[\eta(x+h) - \eta(x)] - \eta'(x)$$

$$= \frac{1}{h}\int_0^h [\eta'(x+t) - \eta'(x)]\,dt \to 0, \quad h \to 0,$$

uniformly for all $(x, u) \in G$. This now implies that the two conditions $\Delta \to 0$, $h \to 0$, and $\varphi \to f$, $h \to 0$, are equivalent. □

Theorem 10.18 *The Euler method is consistent. If f is continuously differentiable in G, then the Euler method has consistency order one.*

Proof. Consistency is a consequence of Theorem 10.17 and the fact that $\varphi(x, u; h) = f(x, u)$ for Euler's method. If f is continuously differentiable, then from the differential equation $\eta' = f(\xi, \eta)$ it follows that η is twice continuously differentiable with

$$\eta'' = f_x(\xi, \eta) + f_u(\xi, \eta) f(\xi, \eta). \tag{10.12}$$

Therefore, Taylor's formula yields

$$|\Delta(x, u; h)| = \left|\frac{1}{h}\,[\eta(x+h) - \eta(x)] - \eta'(x)\right| = \frac{h}{2}\,|\eta''(x + \theta h)| \le Kh$$

for some $0 < \theta < 1$ and some bound K for the function $2(f_x + f_u f)$. □

Theorem 10.19 *The improved Euler method is consistent. If f is twice continuously differentiable in G, then the improved Euler method has consistency order two.*

Proof. Consistency follows from Theorem 10.17 and

$$\varphi(x,u;h) = \frac{1}{2}\,[f(x,u) + f(x+h,u+hf(x,u))] \to f(x,u), \quad h \to 0.$$

If f is twice continuously differentiable, then (10.12) implies that η is three times continuously differentiable with

$$\eta''' = f_{xx}(\xi,\eta) + 2f_{xu}(\xi,\eta)f(\xi,\eta) + f_{uu}(\xi,\eta)f^2(\xi,\eta)$$
$$+f_u(\xi,\eta)f_x(\xi,\eta) + f_u^2(\xi,\eta)f(\xi,\eta).$$

Hence Taylor's formula yields

$$\left|\eta(x+h) - \eta(x) - h\eta'(x) - \frac{h^2}{2}\,\eta''(x)\right| = \frac{h^3}{6}\,|\eta'''(x+\theta h)| \le K_1 h^3 \quad (10.13)$$

for some $0 < \theta < 1$ and a bound K_1 for $6(f_{xx}+2f_{xu}f+f_{uu}f^2+f_uf_x+f_u^2f)$. From Taylor's formula for functions of two variables we have the estimate

$$|f(x+h,u+k) - f(x,u) - hf_x(x,u) - kf_u(x,u)| \le \frac{1}{2}\,K_2(|h|+|k|)^2$$

with a bound K_2 for the second derivatives f_{xx}, f_{xu}, and f_{uu}. From this, setting $k = hf(x,u)$, in view of (10.12) we obtain

$$|f(x+h,u+hf(x,u)) - f(x,u) - h\eta''(x)| \le \frac{1}{2}\,K_2(1+K_0)^2h^2$$

with some bound K_0 for f, whence

$$\left|\varphi(x,u;h) - f(x,u) - \frac{h}{2}\,\eta''(x)\right| \le \frac{1}{4}\,K_2(1+K_0)^2h^2 \qquad (10.14)$$

follows. Now combining (10.13) and (10.14), with the aid of the triangle inequality and using the differential equation, we can establish consistency order two. □

We proceed by investigating the convergence of single-step methods as the step size h tends to zero. This is done for the solution to the initial value problem in a fixed interval $[a,b]$ with initial data at $x_0 = a$ and the step size h and the number n of steps chosen such that $x_n = b$.

Definition 10.20 *Assume that in the interval $[a,b]$ at the equidistant grid points*

$$x_j := x_0 + jh, \quad j = 0,1,\dots,n,$$

with $x_0 = a$ and $x_n = b$, approximate values u_j for the solution $u(x_j)$ to the initial value problem

$$u' = f(x, u), \quad u(x_0) = u_0,$$

are obtained by a single-step method. Then

$$e_j = e_j(h) := u_j - u(x_j), \quad j = 0, 1, \dots, n,$$

is called the global error, *and*

$$E = E(h) := \max_{j=0,\dots,n} |e_j(h)|$$

is called the maximal global error. *The single-step method is called* convergent *if*

$$\lim_{h \to 0} E(h) = 0,$$

and it is said to have convergence order p *if*

$$E(h) \le H h^p$$

for all $h > 0$ and some constant H.

The following lemma is needed for our convergence analysis.

Lemma 10.21 *Let (ξ_j) be a sequence in $\mathbb{R}$ with the property*

$$|\xi_{j+1}| \le (1 + A)|\xi_j| + B, \quad j = 0, 1, \dots,$$

for some constants $A > 0$ and $B \ge 0$. Then the estimate

$$|\xi_j| \le |\xi_0| e^{jA} + \frac{B}{A}\left(e^{jA} - 1\right), \quad j = 0, 1, \dots,$$

holds.

Proof. We prove this by induction. The estimate is true for $j = 0$. Assume that it has been proven for some $j \ge 0$. Then, with the aid of the inequality $1+A < e^A$, which follows from the power series for the exponential function, we obtain

$$\begin{aligned} |\xi_{j+1}| &\le (1 + A)|\xi_0| e^{jA} + (1 + A)\frac{B}{A}\left(e^{jA} - 1\right) + B \\ &\le |\xi_0| e^{(j+1)A} + \frac{B}{A}\left(e^{(j+1)A} - 1\right); \end{aligned}$$

i.e., the estimate also holds for $j + 1$. □

Theorem 10.22 *Assume that the function φ describing the single-step method is continuous (also with respect to h) and satisfies a Lipschitz condition; i.e.,*

$$|\varphi(x,u;h) - \varphi(x,v;h)| \leq M|u-v|$$

for all (x,u), $(x,v) \in G$, all (sufficiently small) h, and a Lipschitz constant M. Then the single-step method is convergent if and only if it is consistent.

Proof. We first show that consistency implies convergence and assume that the single-step method is consistent. For the difference of two consecutive errors we compute

$$\begin{aligned} e_{j+1} - e_j &= [u_{j+1} - u_j] - [u(x_{j+1}) - u(x_j)] \\ &= h\varphi(x_j, u_j; h) - [u(x_{j+1}) - u(x_j)] \\ &= h[\varphi(x_j, u_j; h) - \varphi(x_j, u(x_j); h) - \Delta(x_j, u(x_j); h)]. \end{aligned}$$

Hence

$$|e_{j+1} - e_j| \leq h[M|u_j - u(x_j)| + c(h)], \tag{10.15}$$

where

$$c(h) := \max_{a \leq x \leq b} |\Delta(x, u(x); h)|$$

satisfies

$$c(h) \to 0, \; h \to 0,$$

since we assume consistency. The inequality (10.15) implies that

$$|e_{j+1}| \leq (1 + hM)|e_j| + hc(h), \quad j = 0, 1, \ldots, n.$$

From this, applying Lemma 10.21 for $A = hM$ and $B = hc(h)$ and using $e_0 = 0$, we obtain the estimate

$$|e_j| \leq \frac{c(h)}{M} \left(e^{M(x_j - x_0)} - 1\right), \quad j = 0, 1, \ldots, n. \tag{10.16}$$

This establishes the convergence

$$E(h) \leq \frac{c(h)}{M} \left(e^{M(b-a)} - 1\right) \to 0, \quad h \to 0.$$

We now show that convergence implies consistency and assume that the single-step method is convergent; i.e., for $h \to 0$ the approximations

$$u_{j+1} := u_j + h\varphi(x_j, u_j; h) \tag{10.17}$$

converge to the solution of

$$u'(x) = f(x,u), \quad u(x_0) = u_0,$$

for all initial data $(x_0, u_0) \in G$. We set

$$g(x,u) := \varphi(x,u;0)$$

and observe that by Theorem 10.17 the single-step method is also consistent with the initial value problem

$$u'(x) = g(x,u), \quad u(x_0) = u_0. \tag{10.18}$$

Since we have already shown that consistency implies convergence, the approximations (10.17) also converge to the solution of (10.18); i.e., the solutions of the two initial value problems coincide. Therefore, we have $f(x_0, u_0) = g(x_0, u_0)$, and since this holds for all $(x_0, u_0) \in G$, from the continuity of φ we conclude uniform convergence:

$$\varphi(x,u;h) \to f(x,u), \quad h \to 0.$$

Now consistency follows from Theorem 10.17. □

Theorem 10.23 *Assume that the single-step method satisfies the assumptions of the previous Theorem 10.22 and that it has consistency order p; i.e.,* $|\Delta(x,u;h)| \leq Kh^p$. *Then*

$$|e_j| \leq \frac{K}{M}\left(e^{M(x_j - x_0)} - 1\right) h^p, \quad j = 0, 1, \dots, n;$$

i.e., the convergence also has order p.

Proof. This follows from (10.16) with the aid of $c(h) \leq Kh^p$. □

Corollary 10.24 *The Euler method and the improved Euler method are convergent. For continuously differentiable f the Euler method has convergence order one. For twice continuously differentiable f the improved Euler method has convergence order two.*

Proof. By Theorems 10.18, 10.19, 10.22, and 10.23 it remains only to verify the Lipschitz condition of the function φ for the improved Euler method given by (10.11). From the Lipschitz condition for f we obtain

$$|\varphi(x,u;h) - \varphi(x,v;h)|$$

$$\leq \frac{1}{2}|f(x,u) - f(x,v)| + \frac{1}{2}|f(x+h, u+hf(x,u)) - f(x+h, v+hf(x,v))|$$

$$\leq \frac{L}{2}|u - v| + \frac{L}{2}|[u + hf(x,u)] - [v + hf(x,v)]| \leq L\left(1 + \frac{hL}{2}\right)|u - v|;$$

i.e., φ also satisfies a Lipschitz condition. □

Single-step methods of higher order can be constructed as follows. For a set of real numbers s_ℓ, $\ell = 2, \dots, m$, $c_{\ell i}$, $i = 1, \dots, \ell - 1$, $\ell = 2, \dots, m$, and α_ℓ, $\ell = 1, \dots, m$, the quantities

$$
\begin{aligned}
k_1 &= f(x_j, u_j), \\
k_2 &= f(x_j + s_2 h, u_j + c_{21} k_1 h), \\
k_3 &= f(x_j + s_3 h, u_j + c_{31} k_1 h + c_{32} k_2 h), \\
&\;\;\vdots \\
k_m &= f\left(x_j + s_m h, u_j + h \sum_{i=1}^{m-1} c_{mi} k_i\right)
\end{aligned}
$$

are computed recursively, and then the approximation is obtained by

$$
u_{j+1} = u_j + h \sum_{i=1}^{m} \alpha_i k_i .
$$

The Euler method is described by $m = 1$ and $\alpha_1 = 1$ and the improved Euler method by $m = 2$, $s_2 = 1$, $c_{21} = 1$, and $\alpha_1 = \alpha_2 = 1/2$. The basic goal in the design of higher-order methods is, for a given m, to determine the coefficients such that the order of consistency and hence the order of convergence becomes as large as possible. As an example, we shall consider the Runge–Kutta method, which is the most widely used and most successful single-step method. It was introduced by Runge in 1895 for a single differential equation and extended to systems of differential equations by Kutta in 1901.

Definition 10.25 *The* Runge–Kutta method *for the numerical solution of the initial value problem (10.7) constructs approximations u_j to the exact solution $u(x_j)$ at the equidistant grid points*

$$
x_j := x_0 + jh, \quad j = 1, 2, \dots,
$$

with step size h by using the above higher-order method with

$$
\begin{aligned}
k_1 &= f(x_j, u_j), \\
k_2 &= f\left(x_j + \frac{h}{2}, u_j + \frac{h}{2} k_1\right), \\
k_3 &= f\left(x_j + \frac{h}{2}, u_j + \frac{h}{2} k_2\right), \\
k_4 &= f(x_j + h, u_j + h k_3),
\end{aligned}
$$

and

$$u_{j+1} = u_j + \frac{h}{6}\,(k_1 + 2k_2 + 2k_3 + k_4).$$

For the differential equation $u' = f(x)$ the Runge–Kutta method coincides with Simpson's rule for numerical integration.

Theorem 10.26 *The Runge–Kutta method is consistent. If f is four-times continuously differentiable, then it has consistency order four and hence convergence order four.*

Proof. The function φ describing the Runge–Kutta method is given recursively by

$$\varphi = \frac{1}{6}\,(\varphi_1 + 2\varphi_2 + 2\varphi_3 + \varphi_4),$$

where

$$\varphi_1(x,u;h) = f(x,u),$$

$$\varphi_2(x,u;h) = f\left(x + \frac{h}{2}\,, u + \frac{h}{2}\,\varphi_1(x,u;h)\right),$$

$$\varphi_3(x,u;h) = f\left(x + \frac{h}{2}\,, u + \frac{h}{2}\,\varphi_2(x,u;h)\right),$$

$$\varphi_4(x,u;h) = f(x+h, u + h\varphi_3(x,u;h)).$$

From this, consistency follows immediately by Theorem 10.17.

Analogously to the proof of Theorem 10.18 for the improved Euler method, the consistency order four can be established by a Taylor expansion of $\varphi(x,u;h)$ with respect to powers of h up to order h^4 and expressing the derivatives of η on the right-hand side of

$$\frac{1}{h}\,[\eta(x+h) - \eta(x)] = \eta'(x) + \frac{h}{2}\,\eta''(x) + \frac{h^2}{6}\,\eta'''(x) + \frac{h^3}{24}\,\eta''''(x) + O(h^4)$$

through f and its derivatives by using the differential equation. We leave the details as an exercise for the reader (see Problem 10.9). □

The error estimate in Theorem 10.23 is not practical in general, since the constants M and K have to be determined from higher-order derivatives of f. Therefore, in practice, the error is estimated by the following heuristic consideration. For convergence order p, the error between the approximate solution $\tilde{u}(x;h)$ at the point x, obtained with step size h, and the exact solution $u(x)$ satisfies

$$\tilde{u}(x;h) - u(x) \approx ch^p$$

for some constant c. Correspondingly, for step size $h/2$ we have that

$$\tilde{u}\left(x;\frac{h}{2}\right) - u(x) \approx c\left(\frac{h}{2}\right)^p.$$

Subtracting these relations yields

$$\tilde{u}(x;h) - \tilde{u}\left(x;\frac{h}{2}\right) \approx c\left(\frac{h}{2}\right)^p (2^p - 1).$$

Now the constant c can be eliminated from the last two relations, with the result that

$$\tilde{u}\left(x;\frac{h}{2}\right) - u(x) \approx \frac{1}{2^p - 1}\left[\tilde{u}(x;h) - \tilde{u}\left(x;\frac{h}{2}\right)\right]. \tag{10.19}$$

Hence we may consider (10.19) as an estimate for the error occurring with the smaller step size $h/2$. However, we need to keep in mind that (10.19) does not provide an exact bound and might fail in particular situations. Nevertheless, it can be used for controlling the step size during the course of the numerical calculations in order to adjust the actual step size according to the required accuracy.

Solving for $u(x)$ in (10.19) yields

$$u(x) \approx \frac{2^p\tilde{u}\left(x;\frac{h}{2}\right) - \tilde{u}(x;h)}{2^p - 1}. \tag{10.20}$$

We leave it as an exercise for the reader to interpret (10.20) as a Richardson extrapolation, which we explained in detail for the case of numerical integration in Section 9.5.

10.4 Multistep Methods

In the single-step methods each computed function value of f is used only in one step. It is natural to try to design methods where each computed function value of f is used in several steps. This leads to multistep methods, as described in the following definition.

Definition 10.27 Multistep methods *for the approximate solution of the initial value problem*

$$u' = f(x,u), \quad u(x_0) = u_0,$$

construct approximations u_j *to the exact solution* $u(x_j)$ *at the equidistant grid points*

$$x_j := x_0 + jh, \quad j = 1, 2, \ldots,$$

with step size h *by*

$$u_{j+r} + a_{r-1}u_{j+r-1} + \cdots + a_0 u_j = h\varphi(x_j, u_j, \ldots, u_{j+r-1}; h)$$

for $j = 0, 1, \ldots$. *Here* φ *is a function of* $r + 2$ *variables given in terms of* f, *and* $a_0, \ldots, a_{r-1}$ *are constants.*

To start such a multistep method involving r steps, r starting values $u_0, u_1, \ldots, u_{r-1}$ are required. For example, these can be approximately computed from the initial value u_0 by a single-step method such as the Runge–Kutta method.

A particular class of multistep methods is obtained by approximating the integral in

$$u(x_{j+r}) - u(x_{j+r-k}) = \int_{x_{j+r-k}}^{x_{j+r}} f(\xi, u(\xi))\, d\xi$$

with $1 \leq k \leq r$ by an interpolatory quadrature, i.e., by

$$\int_{x_{j+r-k}}^{x_{j+r}} f(\xi, u(\xi))\, d\xi \approx \int_{x_{j+r-k}}^{x_{j+r}} p(\xi)\, d\xi,$$

where $p \in P_s$ with $0 \leq s \leq r$ is the uniquely determined polynomial with the interpolation property

$$p(x_{j+m}) = f(x_{j+m}, u_{j+m}), \quad m = 0, \ldots, s,$$

i.e., by setting

$$u_{j+r} - u_{j+r-k} = \int_{x_{j+r-k}}^{x_{j+r}} p(\xi)\, d\xi. \tag{10.21}$$

Integrating the Lagrange representation (8.2) of the interpolation polynomial shows that these multistep methods are of the form

$$u_{j+r} - u_{j+r-k} = h \sum_{m=0}^{s} b_m f(x_{j+m}, u_{j+m})$$

with coefficients $b_0, \ldots, b_s$ depending on r, k, and s.

From (10.21) we can generate a variety of methods by choosing the number of steps r, the number $s+1$ of interpolation points, and the number k of integration intervals appropriately. We briefly report on some of these methods.

The *Adams–Bashforth method*, introduced by Adams and Bashforth in 1883, is obtained by taking $k = 1$ and $s = r - 1$. For $r = 1$ the interpolation polynomial is a constant, and therefore

$$u_{j+1} = u_j + h f(x_j, u_j). \tag{10.22}$$

For $r = 2$ the interpolation is linear and leads to

$$u_{j+2} = u_{j+1} + \frac{h}{2}\left[3 f(x_{j+1}, u_{j+1}) - f(x_j, u_j)\right] \tag{10.23}$$

(see Problem 10.12). The Adams–Bashforth method is explicit. Clearly. (10.22) coincides with the Euler method from Definition 10.8.

The *Adams–Moulton method*, devised by Moulton during World War I, is given by $k = 1$ and $s = r$. For $r = 1$ the interpolation is linear, whence

$$u_{j+1} = u_j + \frac{h}{2}\,[f(x_{j+1}, u_{j+1}) + f(x_j, u_j)]. \tag{10.24}$$

For $r = 2$ the interpolation is quadratic, leading to

$$u_{j+2} = u_{j+1} + \frac{h}{12}\,[5f(x_{j+2}, u_{j+2}) + 8f(x_{j+1}, u_{j+1}) - f(x_j, u_j)] \tag{10.25}$$

(see Problem 10.12). The Adams–Moulton method is implicit. One iteration step for the solution of the nonlinear equation for u_{j+r} starting with the approximation given by the corresponding Adams–Bashforth method leads to a predictor corrector method. Clearly, (10.24) coincides with the implicit Euler method from Definition 10.10.

The explicit method for $k = 2$ and $s = r - 1$ is known as the *Nyström method*, and the implicit method for $k = 2$ and $s = r$ is called the *Milne–Thomson method* (see Problem 10.14).

Definition 10.28 *For each $(x, u) \in G$ denote by $\eta = \eta(\xi)$ the unique solution to the initial value problem*

$$\eta' = f(\xi, \eta), \quad \eta(x) = u,$$

for the initial data (x, u). Then

$$\Delta(x, u; h) := \frac{1}{h}\left[\eta(x + rh) + \sum_{m=0}^{r-1} a_m \eta(x + mh)\right]$$
$$-\varphi(x, \eta(x), \ldots, \eta(x + (r-1)h); h)$$

is called the local discretization error. *The multistep method is called* consistent *(with the initial value problem) if*

$$\lim_{h \to 0} \Delta(x, u; h) = 0$$

uniformly for all $(x, u) \in G$, and it is said to have consistency order p *if*

$$|\Delta(x, u; h)| \le K h^p$$

for all $(x, u) \in G$, all $h > 0$, and some constant K.

Theorem 10.29 *If f is $(s + 1)$-times continuously differentiable, then the multistep methods (10.21) are consistent of order $s + 1$.*

Proof. By construction we have that

$$\Delta(x, u; h) = \frac{1}{h}\int_{x+(r-k)h}^{x+rh} [f(\xi, u(\xi)) - p(\xi)]\, d\xi,$$

where p denotes the polynomial satisfying the interpolation condition

$$p(x + mh) = f(x + mh, \eta(x + mh)), \quad m = 0, \dots, s.$$

By Theorem 8.10 on the remainder in polynomial interpolation, we can estimate

$$|f(\xi, \eta(\xi)) - p(\xi)| \le K h^{s+1}$$

for all ξ in the interval $x + (r - k)h \le \xi \le x + rh$ and some constant K depending on f and its derivatives up to order $s + 1$. □

Analyzing the convergence for multistep methods is more involved than for single-step methods for the following two reasons. Firstly, the approximation obtained by a multistep method is, of course, also influenced by the errors

$$e_j := u_j - u(x_j), \quad j = 0, \dots, r - 1,$$

in the starting values. Hence we give the following definition.

Definition 10.30 *The starting values u_j, $j = 0, \dots, r - 1$, are called* consistent *if*

$$\lim_{h \to 0} [u_j(h) - u(x_j)] = 0, \quad j = 0, \dots, r - 1.$$

They are said to have consistency order p *if*

$$|u_j(h) - u(x_j)| \le K^* h^p, \quad j = 0, \dots, r - 1,$$

for all $h > 0$ and some constant K^*.

To make sure that the consistency order of the starting values coincides with the consistency order of the multistep method, the single-step method for computing the starting values has to be chosen accordingly.

Secondly, multistep methods can be unstable, as illustrated by the following example.

Example 10.31 Let p be the quadratic interpolation polynomial satisfying

$$p(x_j) = u(x_j), \quad j = 0, 1, 2,$$

and approximate

$$u'(x_0) \approx p'(x_0).$$

Using the fact that the approximation for the derivative is exact for polynomials of degree less than or equal to two, simple calculations show that (see Problem 10.15)

$$p'(x_0) = \frac{1}{2h} \left[-u(x_2) + 4u(x_1) - 3u(x_0)\right]. \tag{10.26}$$

If u is three times continuously differentiable, by Theorem 8.10 we have

$$\left|\frac{u(x)-p(x)}{x-x_0}\right| \leq \frac{1}{6}\,\|u'''\|_\infty |(x-x_1)(x-x_2)|,$$

and from this, passing to the limit $x \to x_0$, it follows that the error for the derivative can be estimated by

$$|u'(x_0) - p'(x_0)| \leq \frac{h^2}{3}\,\|u'''\|_\infty. \tag{10.27}$$

By approximating

$$p'(x_0) \approx u'(x_0) = f(x_0, u_0)$$

we derive a multistep method of the form

$$u_{j+2} - 4u_{j+1} + 3u_j = -2hf(x_j, u_j), \quad j = 0, 1, \ldots. \tag{10.28}$$

From (10.27) it follows that (10.28) is consistent with order two if f is twice continuously differentiable.

Now we consider the initial value problem

$$u' = -u, \quad u(0) = 1,$$

with the solution $u(x) = e^{-x}$. Here the multistep method (10.28) reads

$$u_{j+2} - 4u_{j+1} + (3-2h)u_j = 0, \quad j = 0, 1, \ldots. \tag{10.29}$$

Table 10.3 gives the error $e_j = u_j - e^{-x_j}$ between the approximate and exact solutions for the step sizes $h = 0.1$ and $h = 0.01$. For the starting values, $u_0 = 1$ and $u_1 = e^{-h}$ have been used with ten-decimal-digits accuracy. The last column gives the quotient $q_j := e_j/e_{j-1}$ of the error in two consecutive steps.

TABLE 10.3. Numerical results for Example 10.31

$h = 0.1$			
j	x_j	e_j	q_j
4	0.4	0.0109	3.60
6	0.6	0.1123	3.15
8	0.8	1.0858	3.10
10	1.0	10.4143	3.10

$h = 0.01$			
j	x_j	e_j	q_j
5	0.05	0.0000	3.23
10	0.10	0.0099	3.01
15	0.15	2.4456	3.01
20	0.20	604.1985	3.01

In order to explain the numerical failure indicated by the results in Table 10.3, we solve the difference equation (10.29) by looking for solutions of the form

$$u_j = a\lambda^j, \tag{10.30}$$

where a and λ are complex numbers. Substituting into (10.29) shows that (10.30) solves (10.29) if and only if λ is a solution of the so-called characteristic equation

$$\lambda^2 - 4\lambda + (3 - 2h) = 0.$$

This quadratic equation has two solutions, namely

$$\lambda_{1,2} = 2 \mp \sqrt{1 + 2h}.$$

Therefore, the general solution of (10.29) is given by

$$u_j = a\lambda_1^j + b\lambda_2^j.$$

The two constants a and b are determined by the conditions $u_0 = 1$ and $u_1 = e^{-h}$ and have the values

$$a = \frac{\lambda_2 - e^{-h}}{\lambda_2 - \lambda_1} = 1 + O(h^2)$$

and

$$b = \frac{e^{-h} - \lambda_1}{\lambda_2 - \lambda_1}.$$

The term $a\lambda_1^j$ in the solution to the difference equation approximates the solution $e^{-x_j} = e^{-jh}$ to the initial value problem, since

$$a\lambda_1^j = [1 + O(h^2)]\,[1 - h + O(h^2)]^j \approx e^{-jh}.$$

However, the additional term $b\lambda_2^j$ grows exponentially, and the relation

$$\frac{u_j - u(x_j)}{u_{j-1} - u(x_{j-1})} \approx \lambda_2 = 3 + h + O(h^2)$$

explains the last column of Table 10.3. □

Roughly speaking, for multistep methods with $r \geq 2$, the (homogeneous) difference equation of order r occurring in the multistep method has r linearly independent solutions, whereas the approximated differential equation has only one solution. Hence only one of the solutions to the difference equation corresponds to the differential equation. Therefore, convergence of the multistep method can be expected only when the additional solutions to the difference equation remain bounded. Note that these additional solutions will always be activated by errors in the starting values and by round-off errors. For this reason we proceed by investigating the stability of the difference equation.

Definition 10.32 *The* linear difference equation

$$u_{j+r} + \sum_{m=0}^{r-1} a_m u_{j+m} = 0, \quad j = 0, 1, \dots, \tag{10.31}$$

with constant coefficients $a_0, \dots, a_{r-1}$ *is called* stable *if all its solutions are bounded.*

Theorem 10.33 *The linear difference equation (10.31) is stable if and only if it satisfies the* root condition, *i.e., if all the zeros λ of the* characteristic polynomial

$$p(\lambda) := \lambda^r + \sum_{m=0}^{r-1} a_m \lambda^m \tag{10.32}$$

have absolute value $|\lambda| \le 1$, and zeros satisfying $|\lambda| = 1$ are simple zeros.

Proof. We begin by noting that each solution to the difference equation (10.31) is uniquely determined by its r initial values $u_0, u_1, \dots, u_{r-1}$. Obviously, from these initial values the remaining terms $u_r, u_{r+1}, \dots$ are recursively determined by (10.31).

For convenience we set $a_r = 1$ and denote by A the differential operator given by $(Af)(\lambda) = \lambda f'(\lambda)$. Then for the sequence

$$u_j = j^n \lambda^j, \quad j = 0, 1, \dots, \tag{10.33}$$

we have that

$$\begin{aligned} u_{j+r} + \sum_{m=0}^{r-1} a_m u_{j+m} &= \sum_{m=0}^{r} a_m (j+m)^n \lambda^{j+m} \\ &= \lambda^j \sum_{k=0}^{n} \binom{n}{k} j^k \sum_{m=0}^{r} a_m\, m^{n-k} \lambda^m \\ &= \lambda^j \sum_{k=0}^{n} \binom{n}{k} j^k (A^{n-k}p)(\lambda). \end{aligned}$$

From this it can be deduced that if λ is a zero of the characteristic polynomial p of multiplicity s, then for $n = 0, 1, \dots, s-1$ the sequence (10.33) solves the difference equation.

Now assume that $\lambda_1, \dots, \lambda_k$ are the zeros of the characteristic polynomial (10.32) and have multiplicities $s_1, \dots, s_k$; i.e.,

$$p(\lambda) = \prod_{l=1}^{k} (\lambda - \lambda_l)^{s_l}.$$

Then the general solution of the homogeneous difference equation (10.31) is given by

$$u_j = \sum_{l=1}^{k} \sum_{s=0}^{s_l - 1} \alpha_{ls} j^s \lambda_l^j \tag{10.34}$$

with r arbitrary constants α_{ls}. To establish this we need to show that the coefficients α_{ls} can be chosen such that arbitrarily given initial conditions

$$\sum_{l=1}^{k} \sum_{s=0}^{s_l - 1} \alpha_{ls} j^s \lambda_l^j = u_j, \quad j = 0, \dots, r-1, \tag{10.35}$$

are fulfilled. The homogeneous adjoint system to the system (10.35) reads

$$\sum_{j=0}^{r-1} \beta_j j^s \lambda_l^j = 0, \quad s = 0, \dots, s_l - 1,\ l = 1, \dots, k.$$

Assume that β_j, $j = 0, \dots, r-1$ is a solution. Then the polynomial

$$q(\lambda) := \sum_{j=0}^{r-1} \beta_j \lambda^j$$

of degree $r-1$ has the zeros λ_l with multiplicity s_l for $l = 1, \dots, k$; i.e., the polynomial has r zeros and therefore, by Theorem 8.1, must vanish identically. This implies $\beta_0 = \cdots = \beta_{r-1} = 0$. Hence, for each given right-hand side the system (10.35) has a unique solution.

Now from the form (10.34) of the general solution to the difference equation, the equivalence of stability and the root condition is obvious. □

Besides the solution (10.34) of the homogeneous difference equation, we also will need an explicit expression for the solution to the inhomogeneous difference equation.

Lemma 10.34 *For $k = 0, 1, \dots, r-1$, let $u_{j,k}$ denote the unique solutions to the homogeneous difference equation (10.31) with initial values*

$$u_{j,k} = \delta_{j,k}, \quad j = 0, 1, \dots, r-1.$$

Then for a given right-hand side $c_r, c_{r+1}, \dots$, the unique solution to the inhomogeneous difference equation

$$z_{j+r} + \sum_{m=0}^{r-1} a_m z_{j+m} = c_{j+r}, \quad j = 0, 1, \dots, \tag{10.36}$$

with initial values $z_0, z_1, \dots, z_{r-1}$ is given by

$$z_{j+r} = \sum_{k=0}^{r-1} z_k\, u_{j+r,k} + \sum_{k=0}^{j} c_{k+r}\, u_{j+r-k-1,r-1}, \quad j = 0, 1, \dots. \tag{10.37}$$

Proof. Setting $u_{m,r-1} = 0$ for $m = -1, -2, \dots$, we can rewrite (10.37) in the form

$$z_j = \sum_{k=0}^{r-1} z_k\, u_{j,k} + w_j, \quad j = 0, 1, \dots,$$

where

$$w_j := \sum_{k=0}^{\infty} c_{k+r}\, u_{j-k-1,r-1}, \quad j = 0, 1, \dots.$$

Obviously, $w_j = 0$ for $j = 0, \dots, r-1$, and therefore it remains to show that w_j satisfies the inhomogeneous difference equation (10.36).

As in the proof of Theorem 10.33 we set $a_r = 1$. Then, using $u_{m,r-1} = 0$ for $m < r-1$, $u_{r-1,r-1} = 1$, and the homogeneous difference equation for $u_{m,r-1}$, we compute

$$\sum_{m=0}^{r} a_m w_{j+m} = \sum_{m=0}^{r} a_m \sum_{k=0}^{\infty} c_{k+r} u_{j+m-k-1,r-1}$$

$$= \sum_{m=0}^{r} a_m \sum_{k=0}^{j} c_{k+r} u_{j+m-k-1,r-1}$$

$$= \sum_{k=0}^{j} c_{k+r} \sum_{m=0}^{r} a_m u_{j+m-k-1,r-1} = c_{j+r}.$$

Now the proof is completed by noting that each solution to the inhomogeneous difference equation (10.36) is uniquely determined by its r initial values $z_0, z_1, \dots, z_{r-1}$. □

Definition 10.35 *The multistep method of Definition 10.27 is called* stable *if the associated difference equation*

$$u_{j+r} + \sum_{m=0}^{r-1} a_m u_{j+m} = 0$$

is stable.

Single-step methods are always stable, since the associated difference equation $u_{j+1} - u_j = 0$ clearly satisfies the root condition.

Remark 10.36 *The multistep methods (10.21) are stable.*

Proof. The corresponding characteristic polynomial $p(\lambda) := \lambda^r - \lambda^{r-k}$ fulfills the root condition. □

For establishing convergence of multistep methods, we will need the following extension of Lemma 10.21.

Lemma 10.37 *Let (ξ_j) be a sequence in $\mathbb{R}$ with the property*

$$|\xi_j| \le A \sum_{m=0}^{j-1} |\xi_m| + B, \quad j = 1, 2, \dots,$$

for some constants $A > 0$ and $B \ge 0$. Then the estimate

$$|\xi_j| \le (A|\xi_0| + B)e^{(j-1)A}, \quad j = 1, 2, \dots,$$

holds.

Proof. We prove by induction that

$$|\xi_j| \le (A|\xi_0| + B)(1+A)^{j-1}, \quad j = 1, 2, \dots. \tag{10.38}$$

Then the assertion follows by using the estimate $1 + A \le e^A$. The inequality (10.38) is true for $j = 1$. Assume that it has been proven up to some $j \ge 1$. Then we have

$$|\xi_{j+1}| \le A \sum_{m=0}^{j} |\xi_m| + B \le (A|\xi_0| + B) + A \sum_{m=1}^{j} (A|\xi_0| + B)(1+A)^{m-1}$$

$$= (A|\xi_0| + B)(1+A)^j;$$

i.e., the estimate is also true for $j + 1$. □

Theorem 10.38 *Assume that the function φ describing the multistep method is continuous and satisfies a Lipschitz condition; i.e.,*

$$|\varphi(x, u_0, u_1, \dots, u_{r-1}; h) - \varphi(x, v_0, v_1, \dots, v_{r-1}; h)| \le M \sum_{m=0}^{r-1} |u_m - v_m|$$

for all $(x, u_0), \dots, (x, u_{r-1})(x, v_0), \dots, (x, v_{r-1}) \in G$, all (sufficiently small) h, and a Lipschitz constant M. Furthermore, assume that the multistep method is consistent and stable and that the starting values are consistent. Then the multistep method is convergent. If both the multistep method and the starting values have consistency order p, then the convergence also is of order p.

Proof. (Compare to the proof of Theorem 10.22.) For the errors

$$e_j := u_j - u(x_j)$$

we obtain

$$e_{j+r} + \sum_{m=0}^{r-1} a_m e_{j+m} = u_{j+r} + \sum_{m=0}^{r-1} a_m u_{j+m} - u(x_{j+r}) - \sum_{m=0}^{r-1} a_m u(x_{j+m})$$

$$= h\varphi(x_j, u_j, \dots, u_{j+r-1}; h) - h\Delta(x_j, u(x_j); h)$$

$$-h\varphi(x_j, u(x_j), \dots, u(x_{j+r-1}); h).$$

We rewrite this into the form

$$e_{j+r} + \sum_{m=0}^{r-1} a_m e_{j+m} = hc_{j+r}, \quad j = 0, 1, \dots, \tag{10.39}$$

where

$$c_{j+r} := \varphi(x_j, u_j, \dots, u_{j+r-1}; h) - \Delta(x_j, u(x_j); h)$$

$$-\varphi(x_j, u(x_j), \dots, u(x_{j+r-1}); h).$$

We can estimate the right-hand side by

$$|c_{j+r}| \le M \sum_{m=0}^{r-1} |e_{j+m}| + c(h), \quad j = 0, 1, \dots, \tag{10.40}$$

where

$$c(h) = \max_{a \le x \le b} |\Delta(x, u(x); h)|$$

satisfies $c(h) \to 0$, $h \to 0$, since we assume consistency. By Lemma 10.34 we can express the solution of (10.39) in the form

$$e_{j+r} = \sum_{k=0}^{r-1} e_k \, u_{j+r,k} + h \sum_{k=0}^{j} c_{k+r} \, u_{j+r-k-1,r-1}, \quad j = 0, 1, \dots .$$

From this, since we assume stability, we can estimate

$$|e_{j+r}| \le N \left\{ d(h) + h \sum_{k=0}^{j} |c_{k+r}| \right\}, \quad j = 0, 1, \dots,$$

for some constant N and

$$d(h) := \sum_{k=0}^{r-1} |e_k|.$$

We note that $d(h) \to 0$, $h \to 0$, since the starting values are assumed to be consistent. Inserting (10.40) into the last inequality now yields

$$|e_{j+r}| \le N \left\{ d(h) + hM \sum_{k=0}^{j} \sum_{m=0}^{r-1} |e_{k+m}| + (j+1)hc(h) \right\}, \quad j = 0, 1, \dots .$$

Because of

$$\sum_{k=0}^{j} \sum_{m=0}^{r-1} |e_{k+m}| = \sum_{m=0}^{r-1} \sum_{k=m}^{m+j} |e_k| \le r \sum_{k=0}^{r+j-1} |e_k| = r \sum_{k=r}^{r+j-1} |e_k| + rd(h)$$

and $(j+1)h \le x_{j+1} - x_0 \le 2(b-a)$ we obtain that $|e_r| \le C\gamma(h)$ and

$$|e_{j+r}| \le C \left\{ \gamma(h) + h \sum_{k=r}^{r+j-1} |e_k| \right\}, \quad j = 1, 2, \dots,$$

for some constant C and $\gamma(h) := d(h) + c(h)$. Now Lemma 10.37 implies that

$$|e_{j+r}| \le C[|e_r|h + \gamma(h)]e^{(j-1)Ch} \le C(1+Ch)\gamma(h)e^{(x_j - x_0)C}, \quad j = 1, 2, \dots,$$

whence

$$E(h) \le C(1 + Ch)\gamma(h)e^{(b-a)C} \to 0, \quad h \to 0,$$

follows, since $\gamma(h) \to 0$, $h \to 0$. For consistency order p we have that $\gamma(h) = O(h^p)$; i.e., the convergence is also of order p. □

The basic advantage of multistep methods results from the fact that for arbitrary convergence order, in each step only one new evaluation of the function f is required. In contrast, for single-step methods the number of function evaluations required in each step is equal, in general, to the convergence order. Therefore, multistep methods are much faster than single-step methods. However, it should be noted that readjusting the step size during the computation is more involved due to the need to recompute the corresponding starting values for the new step size.

Problems

10.1 Find the exact solution of the initial value problem

$$u' = -u^2, \quad u(0) = 1,$$

and compare it to the approximate solutions obtained by successive approximations according to Corollary 10.6. Compute the third iterate u_3 and compare the exact error $u - u_3$ to the a posteriori error estimate from Corollary 10.6.

10.2 Consider the initial value problem $u' = u$, $u(0) = 1$, and show that the approximate solution from the Euler method is given by $u_j = (1 + h)^j$.

10.3 Find the exact solution of the initial value problem

$$u' = 2\,\frac{u}{x}, \quad u(1) = 1.$$

Determine an analytic expression for the approximate solution by Euler's method and verify the convergence order one predicted by Theorem 10.18.

10.4 Show that Euler's method fails to approximate the solution $u(x) = \left(\frac{2}{3}\,x\right)^{3/2}$ of the initial value problem $u' = u^{1/3}$, $u(0) = 0$. Explain this failure.

10.5 Show that the differential equation $u' = ax$ with $a \in \mathbb{R}$ is solved exactly by the improved Euler method.

10.6 Show that the single-step method

$$u_{j+1} = u_j + hf\left(x_j + \frac{h}{2}, u_j + \frac{h}{2}\, f(x_j, u_j)\right)$$

has consistency order two if f is twice continuously differentiable. This method is known as the *modified Euler method.*

10.7 Show that the single-step method given by

$$k_1 = f(x_j, u_j),$$

$$k_2 = f\left(x_j + \frac{h}{3}, u_j + \frac{h}{3}\, k_1\right),$$

$$k_3 = f\left(x_j + \frac{2h}{3}, u_j + \frac{2h}{3}\, k_2\right),$$

and

$$u_{j+1} = u_j + \frac{h}{4}\,(k_1 + 3k_3)$$

is consistent and has consistency order three if f is three-times continuously differentiable. This method is known as *Heun's third-order method.*

10.8 Show that the single-step method given by

$$k_1 = f(x_j, u_j),$$

$$k_2 = f\left(x_j + \frac{h}{2}, u_j + \frac{h}{2}\, k_1\right),$$

$$k_3 = f(x_j + h, u_j - hk_1 + 2hk_2),$$

and

$$u_{j+1} = u_j + \frac{h}{6}\,(k_1 + 4k_2 + k_3)$$

is consistent and has consistency order three if f is three-times continuously differentiable. This method is known as *Kutta's third-order method.*

10.9 Show that the Runge–Kutta method (see Definition 10.25) has consistency order four if f is four-times continuously differentiable.

10.10 Write a computer program for the Runge–Kutta method and test it for various examples.

10.11 The population $p = p(t)$ and $q = q(t)$ of two interacting animal species that have a predator prey relationship is modeled by the system of the *Lotka–Volterra equations*

$$\frac{dp}{dt} = \alpha p + \beta pq, \quad \frac{dq}{dt} = \gamma q + \delta pq$$

with constant coefficients $\alpha < 0$, $\beta > 0$, $\gamma > 0$, and $\delta < 0$, complemented by initial conditions $p(0) = p_0$ and $q(0) = q_0$. (Explain the significance of the signs of the constants for the model.) For the coefficients $\alpha = -1$, $\beta = 0.01$, $\gamma = 0.25$, and $\delta = -0.01$, test the stability of the solutions by solving the initial value problem

numerically by the Runge–Kutta method for the four different initial conditions $p_0 = 30 \pm 1$ and $q_0 = 80 \pm 1$. Visualize the numerical results by a phase diagram, i.e., by the curve $\{(p(t), q(t) : t \in [0, T]\}$ for sufficiently large $T > 0$.

10.12 Verify the coefficients in the Adams–Bashforth and Adams–Moulton methods (10.22)–(10.25).

10.13 Determine the coefficients of the Adams–Bashforth and Adams–Moulton methods for $r = 3$.

10.14 The multistep methods (10.21) for $k = 2$ and $s = r - 1$ and for $k = 2$ and $s = r$ are known as the Nyström method and the Milne–Thomson method, respectively. Determine the coefficients of the Nyström method and the Milne–Thomson method for $r = 1$ and $r = 2$.

10.15 Verify the coefficients in the difference formula (10.26).

10.16 Construct a two-step method of the form

$$u_{j+2} + a_1 u_{j+1} + a_0 u_j = h[b_0 f(x_j, u_j) + b_1 f(x_{j+1}, u_{j+1})]$$

that has consistency order two and discuss its stability.

10.17 Find the general solution of the difference equation

$$u_{j+2} - 2a u_{j+1} + a u_j = 1$$

for $0 < a < 1$. Show that $\lim_{j \to \infty} u_j = 1/(1 - a)$.

10.18 Find an explicit expression for the *Fibonacci numbers* a_j, which are defined by $a_0 = a_1 = 1$ and $a_{j+1} = a_j + a_{j-1}$ for $j \geq 1$. Is the root condition of Theorem 10.33 satisfied?

10.19 Attempt to approximate the unique solution $u(x) = 2$ of the initial value problem

$$u' = xu(u - 2), \quad u(0) = 2,$$

numerically by any of the methods described in this chapter. Discuss the results by relating them to the solution of the initial value problem with perturbed initial condition $u(0) = 2 + \alpha$ for small $\alpha \in \mathbb{R}$.

10.20 Consider the approximate solution of the initial value problem

$$u' + 100u = 100, \quad u(0) = 2,$$

by the Euler method. Explain why for an accurate approximation the step size h has to be chosen smaller than $h < 0.02$ despite the fact that the solution is almost constant for x not too small, say, for $x > 0.1$. (This differential equation is an example of a so-called *stiff equation*, for which the numerical solution is rather delicate.)

11
Boundary Value Problems

Whereas in initial value problems the solution is determined by conditions imposed at one point only, boundary value problems for ordinary differential equations are problems in which the solution is required to satisfy conditions at more than one point, usually at the two endpoints of the interval in which the solution is to be found. Since an ordinary differential equation of order n has, in principle, a general solution depending on n parameters, the total number of boundary conditions required to determine a unique solution is n. For an introduction to some of the basic methods for the numerical solution of such boundary value problems we shall confine ourselves to the simplest boundary value problem, which is one for an equation of the second order in which the solution is specified at two distinct points. For more detailed studies we refer to [13, 36, 46].

As opposed to the fundamental Picard–Lindelöf existence and uniqueness theorem for initial value problems, a detailed analysis of the existence and uniqueness theory for nonlinear boundary value problems is more involved and beyond the scope of this introduction. However, for linear boundary value problems the theory is more elementary, and we shall include part of it in our analysis.

For the numerical solution of boundary value problems for ordinary differential equations three different groups of methods are available: shooting methods, finite difference methods, and finite element methods. Whereas shooting methods, which we briefly describe in Section 11.1 and which rely on numerical methods for initial value problems, are restricted to ordinary differential equations, the finite difference and finite element methods can also be applied to boundary value problems for partial differential equa-

tions. Therefore, our presentation of finite difference and finite element methods for linear ordinary differential equations is also meant as a model discussion for the more complicated and more important case of partial differential equations.

Of course, in one chapter only a small part of the theory and the applications of finite difference and finite element methods can be covered. Hence, we set ourselves the task to outline the basic ideas of these methods by considering only the simplest cases. For a solid foundation of the finite element method, we felt it was necessary to include as its theoretical basis a discussion of the Galerkin method for strictly coercive operators, which appears in Section 11.3. This, in turn, made it necessary to present the Lax–Milgram theorem on the existence of solutions for equations with strictly coercive operators.

11.1 Shooting Methods

Consider the boundary value problem for the differential equation of the second order

$$u'' = f(x, u, u'), \quad a \le x \le b, \tag{11.1}$$

with boundary conditions

$$u(a) = \alpha, \quad u(b) = \beta. \tag{11.2}$$

For the sake of simplicity we assume that the function f is defined on $[a, b] \times \mathbb{R}^2$.

Shooting methods attempt to employ the numerical methods described in the previous chapter for initial value problems where, roughly speaking, the initial conditions at $x = a$ are adjusted so that the solution satisfies the required boundary conditions (11.2). For this, in addition to the boundary value problem, we also consider the initial value problem

$$u'' = f(x, u, u'), \quad u(a) = \alpha, \quad u'(a) = s, \tag{11.3}$$

with a real parameter s. Geometrically speaking, the parameter s prescribes the initial slope of the solution curve.

If we assume that f is continuous and satisfies a Lipschitz condition with respect to u and u', then by the Picard–Lindelöf Theorem 10.5, for each $s \in \mathbb{R}$ there exists a unique solution $u(\cdot, s)$ of the initial value problem (11.3). To arrive at a solution to the boundary value problem (11.1)–(11.2), the parameter s has to be chosen such that $u(b, s) = \beta$; i.e., we have to solve the equation

$$F(s) = 0,$$

where the function $F : \mathbb{R} \to \mathbb{R}$ is defined by

$$F(s) := u(b, s) - \beta.$$

For each s the value $F(s)$ can be computed approximately by one of the numerical methods of Chapter 10 for the solution of initial value problems, extended appropriately to the case of a second-order equation. Note that for a nonlinear differential equation the equation $F(s) = 0$ is nonlinear.

For finding a zero of F the Newton method of Section 6.2 can be employed. For the computation of the derivative $F'(s)$, which is required for Newton's method, we assume that the solution u to the initial value problem (11.3) depends in a continuously differentiable manner on the parameter s. This can be assured by appropriate assumptions on f (see [12]). We set

$$v := \frac{\partial u}{\partial s}$$

and differentiate the differential equation and the initial condition (11.3) with respect to s to obtain

$$\begin{aligned} v''(x,s) &= f_u(x,u(x,s),u'(x,s))v(x,s) \\ &\quad + f_{u'}(x,u(x,s),u'(x,s))v'(x,s) \end{aligned} \tag{11.4}$$

and

$$v(a,s) = 0, \quad v'(a,s) = 1. \tag{11.5}$$

Since

$$F'(s) = v(b,s),$$

computing the derivative of F requires solving the additional linear initial value problem (11.4)–(11.5) for v, where u is known from solving (11.3). Note that from a numerical approximation, u is known only at grid points. Summarizing, we obtain the following method.

Algorithm 11.1 *The* shooting method *with* Newton iterations *consists of the following steps:*

1. *Choose an initial slope* $s \in \mathbb{R}$.

2. *Solve numerically the initial value problem for*

$$u'' = f(x,u,u')$$

with initial conditions $u(a) = \alpha$, $u'(a) = s$ *and the initial value problem for*

$$v'' = f_u(x,u,u')v + f_{u'}(x,u,u')v'$$

with initial conditions $v(a) = 0$, $v'(a) = 1$.

3. *If* $u(b) = \beta$ *is satisfied within the required accuracy, then stop; otherwise, replace* s *by*

$$s - \frac{u(b) - \beta}{v(b)}$$

and go back to step 2.

Example 11.2 Consider the boundary value problem

$$u'' = u^3, \quad u(1) = \sqrt{2}, \quad u(2) = \frac{1}{2}\sqrt{2},$$

with the exact solution $u(x) = \sqrt{2}/x$. We solve numerically the associated initial value problem

$$u'' = u^3, \quad u(1) = \sqrt{2}, \quad u'(1) = s,$$

by the improved Euler method of Section 10.2 with step sizes $h = 0.1$, $h = 0.01$, and $h = 0.001$. For this we transform the initial value problem for the equation of second order into the initial value problem for the system

$$u' = w, \quad w' = u^3, \quad u(1) = \sqrt{2}, \quad w(1) = s.$$

As starting value for the Newton iteration we choose $s = 0$. The exact initial condition is $s = -\sqrt{2} = -1.414214$. The numerical results represented in Table 11.1 illustrate the feasibility of the shooting method with Newton iterations. □

TABLE 11.1. Numerical results for Example 11.2.

$h = 0.1$		$h = 0.01$		$h = 0.001$	
s	$F(s)$	s	$F(s)$	s	$F(s)$
0.00000	3.61648	0.00000	3.84079	0.00000	3.84400
−0.81116	1.10056	−0.74681	1.26284	−0.74584	1.26538
−1.31684	0.15879	−1.28234	0.21124	−1.28180	0.21210
−1.41553	0.00373	−1.40987	0.00678	−1.40980	0.00684
−1.41796	0.00000	−1.41424	0.00000	−1.41420	0.00000
−1.41796	0.00000	−1.41424	0.00000	−1.41421	0.00000

Numerical problems with ill-conditioning will arise in cases where small changes in the initial data s will cause large changes in the solution $u(\cdot, s)$. This is illustrated by the following example.

Example 11.3 The linear boundary value problem

$$u'' - u' - 110u = 0, \quad u(0) = u(10) = 1,$$

has the unique solution

$$u(x) = \frac{1}{e^{110} - e^{-100}} \{(e^{110} - 1)e^{-10x} + (1 - e^{-100})e^{11x}\}.$$

The unique solution to the associated initial value problem with initial conditions $u(0) = 1$ and $u'(0) = s$ is given by

$$u(x) = \frac{11 - s}{21}\, e^{-10x} + \frac{10 + s}{21}\, e^{11x}.$$

Hence, in this case we have

$$F(s) = \frac{11-s}{21}\, e^{-100} + \frac{10+s}{21}\, e^{110} - 1.$$

From $F(s) = 0$ we deduce that the exact initial slope s satisfies

$$-10 < s = -10 + 21\, \frac{e^{-110} - e^{-210}}{1 - e^{-210}} \, .$$

In a numerical computation with ten-decimal-digit accuracy the best approximation $\tilde{s}$ to the exact zero s we can expect is such that

$$-10 \le \tilde{s} \le -10 + 10^{-9}.$$

Within this interval of initial conditions we now have

$$u(10, -10) = e^{-100} \approx 0$$

and

$$u(10, -10 + 10^{-9}) = \frac{21 - 10^{-9}}{21}\, e^{-100} + \frac{10^{-9}}{21}\, e^{110} \approx 2.8 \cdot 10^{37};$$

i.e., small changes in s will cause very large changes in the values of the solution at the other endpoint. Hence, we cannot expect that this boundary value problem can be numerically solved by the simple version of the shooting method. □

This difficulty can be remedied by a *multiple shooting method* as follows. The interval $[a, b]$ is subdivided into n subintervals according to

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b.$$

Then for given vectors $u = (u_0, \dots, u_{n-1})^T$ and $s = (s_0, \dots, s_{n-1})^T$ in $\mathbb{R}^n$ such that $u_0 = \alpha$, for $j = 0, \dots, n-1$ consider the n initial value problems for

$$u'' = f(x, u, u')$$

on the subintervals $[x_j, x_{j+1}]$ with initial conditions

$$u(x_j) = u_j, \quad u'(x_j) = s_j.$$

In order to obtain from this a solution to the differential equation on all of the interval $[a, b]$, the solutions $u(\cdot, u_j, s_j)$ on the subintervals $[x_j, x_{j+1}]$ have to coincide at the grid points $x_1, \dots, x_{n-1}$ together with their first derivatives. Then the differential equation ensures that the function is twice continuously differentiable on $[a, b]$. In addition, the boundary condition

$u(b) = \beta$ must be satisfied. Altogether we have the following $2n-1$ nonlinear equations for the $2n-1$ unknowns $u_1, \dots, u_{n-1}$ and $s_0, \dots, s_{n-1}$:

$$
\begin{aligned}
u(x_{j+1}, u_j, s_j) - u_{j+1} &= 0, \quad j = 0, \dots, n-2, \\
u'(x_{j+1}, u_j, s_j) - s_{j+1} &= 0, \quad j = 0, \dots, n-2, \\
u(x_n, u_{n-1}, s_{n-1}) - \beta &= 0.
\end{aligned}
\tag{11.6}
$$

For the solution of this system Newton's method can again be used. For details we refer to [36, 50].

11.2 Finite Difference Methods

As already indicated in Example 2.1, the basic idea of finite difference methods for the approximate solution of boundary value problems consists in replacing the derivatives in the differential equations by difference quotients. For the sake of simplicity, we confine our presentation to a linear boundary value problem. Without loss of generality we need consider only the homogeneous boundary condition, since inhomogeneous boundary conditions can be dealt with by incorporating them into the right-hand side of the differential equation (see Problem 11.3).

Theorem 11.4 *Assume that $q, r \in C[a,b]$ and $q \geq 0$. Then the boundary value problem for the linear differential equation*

$$-u'' + qu = r \quad \text{on } [a,b] \tag{11.7}$$

with homogeneous boundary conditions

$$u(a) = u(b) = 0 \tag{11.8}$$

has a unique solution $u \in C^2[a,b]$.

Proof. Assume that u_1 and u_2 are two solutions to the boundary value problem. Then the difference $u = u_1 - u_2$ solves the homogeneous boundary value problem

$$-u'' + qu = 0, \quad u(a) = u(b) = 0.$$

By partial integration we obtain

$$\int_a^b ([u']^2 + qu^2)\, dx = \int_a^b (-u'' + qu)u\, dx = 0.$$

This implies $u' = 0$ on $[a,b]$, since $q \geq 0$. Hence u is constant on $[a,b]$, and the boundary conditions finally yield $u = 0$ on $[a,b]$. Therefore, the boundary value problem (11.7)–(11.8) has at most one solution.

The general solution of the linear differential equation (11.7) is given by

$$u = C_1 u_1 + C_2 u_2 + u^*, \tag{11.9}$$

where u_1, u_2 denotes a fundamental system of two linearly independent solutions to the homogeneous differential equation, u^* is a solution to the inhomogeneous differential equation, and C_1 and C_2 are arbitrary constants. This can be seen with the help of the Picard–Lindelöf Theorem 10.1 (see Problem 11.4). The boundary condition (11.8) is satisfied, provided that the constants C_1 and C_2 solve the linear system

$$C_1 u_1(a) + C_2 u_2(a) = -u^*(a),$$

$$C_1 u_1(b) + C_2 u_2(b) = -u^*(b).$$

This system is uniquely solvable. Assume that C_1 and C_2 solve the homogeneous system. Then $u = C_1 u_1 + C_2 u_2$ yields a solution to the homogeneous boundary value problem. Hence $u = 0$, since we have already established uniqueness for the boundary value problem. From this we conclude that $C_1 = C_2 = 0$ because u_1 and u_2 are linearly independent, and the existence proof is complete. □

For the numerical solution, proceeding as in Example 2.1, we choose an equidistant grid

$$x_j = a + jh, \quad j = 0, \dots, n+1,$$

with the step size given by $h = (b-a)/(n+1)$ and $n \in \mathbb{N}$. At the internal grid points x_j, $j = 1, \dots, n$, we replace the differential quotient in the differential equation by the difference quotient

$$u''(x_j) \approx \frac{1}{h^2}\left[u(x_{j+1}) - 2u(x_j) + u(x_{j-1})\right]$$

to obtain the system of equations

$$-\frac{1}{h^2}\left[u_{j-1} - (2 + h^2 q_j) u_j + u_{j+1}\right] = r_j, \quad j = 1, \dots, n, \tag{11.10}$$

for approximate values u_j to the exact solution $u(x_j)$. Here we have set $q_j := q(x_j)$ and $r_j := r(x_j)$. The system has to be complemented by the two boundary conditions

$$u_0 = u_{n+1} = 0. \tag{11.11}$$

For an abbreviated notation we introduce the $n \times n$ tridiagonal matrix

$$A = \frac{1}{h^2}\begin{pmatrix} 2+q_1h^2 & -1 & & & & \\ -1 & 2+q_2h^2 & -1 & & & \\ & -1 & 2+q_3h^2 & -1 & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & -1 & 2+q_{n-1}h^2 & -1 \\ & & & & -1 & 2+q_nh^2 \end{pmatrix}$$

and the vectors $U = (u_1, \dots, u_n)^T$ and $R = (r_1, \dots, r_n)^T$. Then our system of equations, including the boundary conditions, reads

$$AU = R. \tag{11.12}$$

The following two questions have to be answered:

1. Is the system (11.12) uniquely solvable?
2. How large is the error between the approximate solution u_j and the exact solution $u(x_j)$? Do we have convergence of the approximate solution to the exact solution as $h \to 0$?

Theorem 11.5 *For each $h > 0$ the difference equations (11.10)–(11.11) have a unique solution.*

Proof. The tridiagonal matrix A is irreducible and weakly row-diagonally dominant. Hence, by Theorem 4.7, the matrix A is invertible, and the Jacobi iterations converge. □

Recall that for speeding up the convergence of the Jacobi iterations we can use relaxation methods or multigrid methods as discussed in Sections 4.2 and 4.3.

The error and convergence analysis is initiated by first establishing the following two lemmas.

Lemma 11.6 *Denote by A the matrix of the finite difference method for $q \ge 0$ and by A_0 the corresponding matrix for $q = 0$. Then*

$$0 \le A^{-1} \le A_0^{-1};$$

i.e., all components of A^{-1} are nonnegative and smaller than or equal to the corresponding components of A_0^{-1}.

Proof. The columns of the inverse $A^{-1} = (a_1, \dots, a_n)$ satisfy $Aa_j = e_j$ for $j = 1, \dots, n$ with the canonical unit vectors $e_1, \dots, e_n$ in $\mathbb{R}^n$. The Jacobi iterations for the solution of $Az = e_j$ starting with $z_0 = 0$ are given by

$$z_{\nu+1} = -D^{-1}(A_L + A_R)z_\nu + D^{-1}e_j, \quad \nu = 0, 1, \dots,$$

with the usual splitting $A = D + A_L + A_R$ of A into its diagonal, lower, and upper triangular parts. Since the entries of D^{-1} and of $-D^{-1}(A_L + A_R)$ are all nonnegative, it follows that $A^{-1} \ge 0$. Analogously, the iterations

$$z_{\nu+1} = -D_0^{-1}(A_L + A_R)z_\nu + D_0^{-1}e_j, \quad \nu = 0, 1, \dots,$$

yield the columns of A_0^{-1}. Therefore, from $D_0^{-1} \ge D^{-1}$ we conclude that $A_0^{-1} \ge A^{-1}$. □

Lemma 11.7 *Assume that $u \in C^4[a,b]$. Then*

$$\left| u''(x) - \frac{1}{h^2}\,[u(x+h) - 2u(x) + u(x-h)] \right| \le \frac{h^2}{12}\,\|u^{(4)}\|_\infty$$

for all $x \in [a+h, b-h]$.

Proof. By Taylor's formula we have that

$$u(x \pm h) = u(x) \pm hu'(x) + \frac{h^2}{2}\,u''(x) \pm \frac{h^3}{6}\,u'''(x) + \frac{h^4}{24}\,u^{(4)}(x \pm \theta_\pm h)$$

for some $\theta_\pm \in (0,1)$. Adding these two equations gives

$$u(x+h)-2u(x)+u(x-h) = h^2u''(x)+\frac{h^4}{24}\,u^{(4)}(x+\theta_+h)+\frac{h^4}{24}\,u^{(4)}(x-\theta_-h),$$

whence the statement of the lemma follows. □

Theorem 11.8 *Assume that the solution to the boundary value problem (11.7)–(11.8) is four-times continuously differentiable. Then the error of the finite difference approximation can be estimated by*

$$|u(x_j) - u_j| \le \frac{h^2}{96}\,\|u^{(4)}\|_\infty (b-a)^2, \quad j = 1, \dots, n.$$

Proof. By Lemma 11.7, for

$$z_j := u''(x_j) - \frac{1}{h^2}\,[u(x_{j+1}) - 2u(x_j) + u(x_{j-1})]$$

we have the estimate

$$|z_j| \le \frac{h^2}{12}\,\|u^{(4)}\|_\infty, \quad j = 1, \dots, n. \tag{11.13}$$

Since

$$-\frac{1}{h^2}\,[u(x_{j+1})-(2+h^2q_j)u(x_j)+u(x_{j-1})] = -u''(x_j)+q_ju(x_j)+z_j = r_j+z_j,$$

the vector $\tilde{U} = (u(x_1), \dots, u(x_n))^T$ given by the exact solution solves the linear system

$$A\tilde{U} = R + Z,$$

where $Z = (z_1, \dots, z_n)^T$. Therefore,

$$A(\tilde{U} - U) = Z,$$

and from this, using Lemma 11.6 and the estimate (11.13), we obtain

$$|u(x_j) - u_j| \le \|A^{-1}Z\|_\infty \le \frac{h^2}{12}\,\|u^{(4)}\|_\infty \|A_0^{-1}e\|_\infty, \quad j = 1, \dots, n, \tag{11.14}$$

where $e = (1, \ldots, 1)^T$. The boundary value problem

$$-u_0'' = 1, \quad u_0(a) = u_0(b) = 0,$$

has the solution

$$u_0(x) = \frac{1}{2}\,(x-a)(b-x).$$

Since $u_0^{(4)} = 0$, in this case, as a consequence of (11.14) the finite difference approximation coincides with the exact solution; i.e., $e = A_0 U = A_0 \tilde{U}$. Hence,

$$\|A_0^{-1} e\|_\infty \leq \|u_0\|_\infty = \frac{1}{8}\,(b-a)^2, \quad j = 1, \ldots, n.$$

Inserting this into (11.14) completes the proof. □

Theorem 11.8 confirms that as in the case of the initial value problems in Chapter 10, the order of the local discretization error is inherited by the global error. Note that the assumption in Theorem 11.8 on the differentiability of the solution is satisfied if q and r are twice continuously differentiable.

The error estimate in Theorem 11.8 is not practical in general, since it requires a bound on the fourth derivative of the unknown exact solution. Therefore, in practice, analogously to (10.19) the error is estimated from the numerical results for step sizes h and $h/2$. Similarly, as in (10.20), a Richardson extrapolation can be employed to obtain a fourth-order approximation.

Of course, the finite difference approximation can be extended to the general linear ordinary differential equation of second order

$$-u'' + pu' + qu = r$$

by using the approximation

$$u'(x_j) \approx \frac{1}{2h}\,[u(x_{j+1}) - u(x_{j-1})] \tag{11.15}$$

for the first derivative. This approximation again has an error of order $O(h^2)$ (see Problem 11.9). Besides Richardson extrapolation, higher-order approximations can be obtained by using higher-order difference approximations for the derivatives such as

$$\begin{aligned} u''(x) \approx \frac{1}{12h^2}[&-u(x+2h) + 16u(x+h) \\ &-30u(x) + 16u(x-h) - u(x-2h)], \end{aligned} \tag{11.16}$$

which is of order $O(h^4)$, provided that u is six-times continuously differentiable (see Problem 11.9).

We wish also to indicate briefly how the finite difference approximations are applied to boundary value problems for partial differential equations. For this we consider the boundary value problem for

$$-\Delta u + qu = r \quad \text{in } D \tag{11.17}$$

in the unit square $D = (0,1) \times (0,1)$ with boundary condition

$$u = 0 \quad \text{on } \partial D. \tag{11.18}$$

Here Δ denotes the Laplacian

$$\Delta u := \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}\,.$$

Proceeding as in the proof of Theorem 11.4, by partial integration it can be seen that under the assumption $q \geq 0$ this boundary value problem has at most one solution. It is more involved and beyond the scope of this book to establish that a solution exists under proper assumptions on the functions q and r. We refer to [24, 60] and also the remarks at the end of Section 11.4.

As in Example 2.2, we choose an equidistant grid

$$x_{ij} = (ih, jh), \quad i, j = 0, \ldots, n+1,$$

with step size $h = 1/(n+1)$ and $n \in \mathbb{N}$. Then we approximate the Laplacian at the internal grid points by

$$\Delta u(x_{ij}) \approx \frac{1}{h^2}\, \{u(x_{i+1,j}) + u(x_{i-1,j}) + u(x_{i,j+1}) + u(x_{i,j-1}) - 4u(x_{ij})\}$$

and obtain the system of equations

$$\begin{aligned} &\frac{1}{h^2}\,[(4 + h^2 q_{ij})u_{ij} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1}] = r_{ij}, \\ &\qquad\qquad\qquad\qquad\qquad\qquad i, j = 1, \ldots, n, \end{aligned} \tag{11.19}$$

for approximate values u_{ij} to the exact solution $u(x_{ij})$. Here we have set $q_{ij} := q(x_{ij})$ and $r_{ij} := r(x_{ij})$. This system has to be complemented by the boundary conditions

$$\begin{aligned} u_{0,j} &= u_{n+1,j} = 0, \quad j = 0, \ldots, n+1, \\ u_{i,0} &= u_{i,n+1} = 0, \quad i = 1, \ldots, n. \end{aligned} \tag{11.20}$$

We refrain from rewriting the system (11.19)–(11.20) in matrix notation and refer back to Example 2.2. Analogously to Theorem 11.5, it can be seen that the Jacobi iterations converge (and relaxation methods and multigrid methods are applicable). Hence we have the following theorem.

Theorem 11.9 *For each $h > 0$ the difference equations (11.19)–(11.20) have a unique solution.*

From the proof of Lemma 11.6 it can be seen that its statement also holds for the corresponding matrices of the system (11.19)–(11.20). Lemma 11.7 implies that

$$\Big|\Delta u(x_1, x_2) - \frac{1}{h^2}\,[u(x_1+h, x_2) + u(x_1-h, x_2) + u(x_1, x_2+h)$$

$$+u(x_1, x_2-h) - 4u(x_1, x_2)]\Big| \le \frac{h^2}{12}\left[\left\|\frac{\partial^4 u}{\partial x_1^4}\right\|_\infty + \left\|\frac{\partial^4 u}{\partial x_2^4}\right\|_\infty\right],$$

provided that $u \in C^4([0,1]\times[0,1])$. Then we can proceed as in the proof of Theorem 11.8 to derive an error estimate. For this we need to have an estimate on the solution of

$$-\Delta u_0 = 1 \quad \text{in } D, \quad u_0 = 0 \quad \text{on } \partial D. \tag{11.21}$$

Either from an explicit form of the solution obtained by separation of variables or by writing

$$u_0(x) = \frac{1}{4}\,(1-x_1)x_1 + \frac{1}{4}\,(1-x_2)x_2 + v_0(x),$$

where v_0 is a harmonic function, i.e., a solution of $\Delta v_0 = 0$, and employing the maximum minimum principle for harmonic functions (see [39]), it can be seen that $\|u_0\|_\infty \le 1/8$ (see Problem 11.10). Hence we can state the following theorem.

Theorem 11.10 *Assume that the solution to the boundary value problem (11.17)–(11.18) is four-times continuously differentiable. Then the error of the finite difference approximation can be estimated by*

$$|u(x_{ij}) - u_{ij}| \le \frac{h^2}{96}\left[\left\|\frac{\partial^4 u}{\partial x_1^4}\right\|_\infty + \left\|\frac{\partial^4 u}{\partial x_2^4}\right\|_\infty\right], \quad i,j = 1,\dots,n.$$

11.3 The Riesz and Lax–Milgram Theorems

To establish the foundation of finite element methods for boundary value problems we need to extend our tools from functional analysis.

Theorem 11.11 (Riesz) *Let X be a Hilbert space. Then for each bounded linear function $F : X \to \mathbb{C}$ there exists a unique element $f \in X$ such that*

$$F(u) = (u, f) \tag{11.22}$$

for all $u \in X$. The norms of the element f and the linear function F coincide; i.e.,

$$\|f\| = \|F\|. \tag{11.23}$$

Proof. Uniqueness follows from the observation that because of the positive definiteness of the scalar product, $f = 0$ is the only element representing the zero function $F = 0$ in the sense of (11.22). For $F \neq 0$ choose $w \in X$ with $F(w) \neq 0$. Since F is continuous, the nullspace

$$N(F) = \{u \in X : F(u) = 0\}$$

can be seen to be a closed, and consequently, by Remark 3.40, a complete, subspace of the Hilbert space X. By the approximation Theorem 3.52 there exists the best approximation v to w with respect to $N(F)$. By Theorem 3.51 it satisfies $w - v \perp N(F)$. Then for $g := w - v$ we have that

$$(F(g)u - F(u)g, g) = 0, \quad u \in X,$$

since $F(g)u - F(u)g \in N(F)$ for all $u \in X$. Hence,

$$F(u) = \left(u, \frac{\overline{F(g)}\, g}{\|g\|^2}\right)$$

for all $u \in X$, which completes the proof of (11.22).

From (11.22) and the Cauchy–Schwarz inequality we have that

$$|F(u)| \leq \|f\|\,\|u\|, \quad u \in X,$$

whence $\|F\| \leq \|f\|$ follows. On the other hand, inserting f into (11.22) yields

$$\|f\|^2 = F(f) \leq \|F\|\,\|f\|,$$

and therefore $\|f\| \leq \|F\|$. This concludes the proof of the norm equality (11.23). □

Definition 11.12 *A linear operator $A : X \to X$ in a pre-Hilbert space X is called* strictly coercive *if there exists a constant $c > 0$ such that*

$$\operatorname{Re}(Au, u) \geq c\|u\|^2 \tag{11.24}$$

for all $u \in X$.

Theorem 11.13 (Lax–Milgram) *In a Hilbert space X a bounded and strictly coercive linear operator $A : X \to X$ has a bounded inverse $A^{-1} : X \to X$.*

Proof. Using the Cauchy–Schwarz inequality, we can estimate

$$\|Au\|\,\|u\| \geq \operatorname{Re}(Au, u) \geq c\|u\|^2.$$

Hence

$$||Au|| \geq c||u|| \tag{11.25}$$

for all $u \in X$. From (11.25) we observe that $Au = 0$ implies $u = 0$; i.e., A is injective.

Next we show that the range $A(X)$ is closed. Let v be an element of the closure $\overline{A(X)}$ and let (v_n) be a sequence from $A(X)$ with $v_n \to v$, $n \to \infty$. Then we can write $v_n = Au_n$ with some $u_n \in X$, and from (11.25) we find that

$$c||u_n - u_m|| \leq ||v_n - v_m||$$

for all n, $m \in \mathbb{N}$. Therefore, (u_n) is a Cauchy sequence in X and converges: $u_n \to u$, $n \to \infty$, with some $u \in X$. Then $v = Au$, since A is continuous, and $A(X) = \overline{A(X)}$ is proven.

From Remark 3.40 we now have that $A(X)$ is complete. Let $w \in X$ be arbitrary and denote by v its best approximation with respect to $A(X)$, which uniquely exists by Theorem 3.52. Then, by Theorem 3.51, we have $(w - v, u) = 0$ for all $u \in A(X)$. In particular, $(w - v, A(w - v)) = 0$. Hence, from (11.24) we see that $w = v \in A(X)$. Therefore, A is surjective. Finally, the boundedness of the inverse

$$||A^{-1}|| \leq \frac{1}{c} \tag{11.26}$$

is a consequence of (11.25). □

Definition 11.14 *Let X be a complex (or real) linear space. Then a function $S : X \times X \to \mathbb{C}$ (or $\mathbb{R}$) is called* sesquilinear *if it is linear with respect to the first variable and antilinear with respect to the second variable, i.e., if*

$$S(\alpha u + \beta v, w) = \alpha S(u, w) + \beta S(v, w)$$

and

$$S(u, \alpha v + \beta w) = \bar{\alpha} S(u, v) + \bar{\beta} S(u, w)$$

for all $u, v, w \in X$ and $\alpha, \beta \in \mathbb{C}$ (or $\mathbb{R}$). A sesquilinear function on a normed space X is called bounded *if*

$$|S(u, v)| \leq C||u|| \, ||v||$$

for all $u, v \in X$ and some positive constant C. It is called strictly coercive *if*

$$\operatorname{Re} S(u, u) \geq c||u||^2$$

for all $u \in X$ and some positive constant c.

Note that for a real linear space, sesquilinear functions are bilinear functions, i.e., linear with respect to both variables. Each bounded and strictly

coercive linear operator $A : X \to X$ in a pre-Hilbert space defines a bounded and strictly coercive sesquilinear function by

$$S(u,v) := (u, Av), \quad u,v \in X.$$

The converse of this statement is described by the following theorem.

Theorem 11.15 *Let S be a bounded and strictly coercive sesquilinear function on a Hilbert space X. Then there exists a uniquely determined bounded and strictly coercive linear operator $A : X \to X$ such that*

$$S(u,v) = (u, Av)$$

for all $u, v \in X$.

Proof. For each $v \in X$ the mapping $u \mapsto S(u,v)$ clearly defines a bounded linear function on X, since $|S(u,v)| \le C\|u\|\,\|v\|$. By the Riesz Theorem 11.11 we can write $S(u,v) = (u,f)$ for all $u \in X$ and some $f \in X$. Therefore, setting $Av := f$ we define an operator $A : X \to X$ such that $S(u,v) = (u, Av)$ for all $u, v \in X$.

To show that A is linear we observe that

$$(u, \alpha Av + \beta Aw) = \bar{\alpha}(u, Av) + \bar{\beta}(u, Aw) = \bar{\alpha}S(u,v) + \bar{\beta}S(u,w)$$

$$= S(u, \alpha v + \beta w) = (u, A[\alpha v + \beta w])$$

for all $u, v, w \in X$ and all $\alpha, \beta \in \mathbb{C}$. The boundedness of A follows from

$$\|Au\|^2 = (Au, Au) = S(Au, u) \le C\|Au\|\,\|u\|,$$

and the strict coercivity of A is a consequence of the strict coercivity of S.

To show uniqueness of the operator A we suppose that there exist two operators A_1 and A_2 with the property

$$S(u,v) = (u, A_1 v) = (u, A_2 v)$$

for all $u, v \in X$. Then we have $(u, A_1 v - A_2 v) = 0$ for all $u, v \in X$, which implies $A_1 v = A_2 v$ for all $v \in X$ by setting $u = A_1 v - A_2 v$. □

Corollary 11.16 *Let S be a bounded and strictly coercive sesquilinear function and F a bounded linear function on a Hilbert space X. Then there exists a unique $u \in X$ such that*

$$S(v,u) = F(v) \tag{11.27}$$

for all $v \in X$.

Proof. By Theorem 11.15 there exists a uniquely determined bounded and strictly coercive linear operator A such that

$$S(v,u) = (v, Au)$$

for all $u, v \in X$, and by Theorem 11.11 there exists a uniquely determined element f such that

$$F(v) = (v, f)$$

for all $v \in X$. Hence, the equation (11.27) is equivalent to the equation

$$Au = f.$$

However, the latter equation is uniquely solvable as a consequence of the Lax–Milgram Theorem 11.13. □

Since the coercivity constants for A and S coincide, from (11.23) and (11.26) we conclude that

$$\|u\| \leq \frac{1}{c}\,\|F\| \tag{11.28}$$

for the unique solution u of (11.27).

Let $A : X \to X$ be a bounded linear operator. Then, given $f \in X$, solving the equation $Au = f$ obviously is equivalent to finding $u \in X$ such that

$$(v, Au) = (v, f) \tag{11.29}$$

for all $v \in X$. The *Galerkin method*, named after the Russian engineer Galerkin, is based on this observation, and given a finite-dimensional subspace $X_n \subset X$, it approximately solves (11.29) by an element $u_n \in X_n$ such that

$$(v, Au_n) = (v, f) \tag{11.30}$$

for all $v \in X_n$. By Theorems 3.51 and 3.52, the condition (11.30) is equivalent to the fact that the best approximations to Au_n and to f with respect to X_n coincide; i.e.,

$$P_n A u_n = P_n f, \tag{11.31}$$

where P_n denotes the orthogonal projection operator from X onto X_n. The equivalence of (11.30) and (11.31) is the reason why the Galerkin method belongs to the so-called *projection methods*; i.e., the equation to be approximated is projected onto a finite-dimensional subspace.

To analyze the Galerkin method we introduce a finite-dimensional operator $A_n : X_n \to X_n$ by $A_n := P_n A$. Then, by Theorem 3.51, we have

$$(A_n u, u) = (P_n A u, u) = (Au, u) + (P_n A u - Au, u) = (Au, u)$$

for all $u \in X_n$. Hence from the strict coercivity of A we deduce that

$$\operatorname{Re}(A_n u, u) \geq c\|u\|^2$$

for all $u \in X_n$; i.e., $A_n : X_n \to X_n$ is strictly coercive with the same coercitivity constant c as $A : X \to X$. This now can be employed to prove the following theorem.

Theorem 11.17 *For a bounded and strictly coercive linear operator A the Galerkin equations* (11.30) *have a unique solution. It satisfies the error estimate*

$$\|u_n - u\| \le M \inf_{v \in X_n} \|v - u\|, \tag{11.32}$$

where M is some constant depending on A (and not on X_n).

Proof. Since $A_n : X_n \to X_n$ is strictly coercive with coercitivity constant c, by the Lax–Milgram Theorem 11.13 we conclude that A_n is bijective; i.e., the Galerkin equations (11.30) have a unique solution $u_n \in X_n$. The estimate (11.26) applied to the operator A_n implies that

$$\|A_n^{-1}\| \le \frac{1}{c} \,. \tag{11.33}$$

For the error $u_n - u$ between the Galerkin approximation u_n and the exact solution u we can write

$$u_n - u = (A_n^{-1} P_n A - I)u = (A_n^{-1} P_n A - I)(u - v)$$

for all $v \in X_n$, since, trivially, we have $A_n^{-1} P_n A v = v$ for $v \in X_n$. By Theorem 3.52 we have $\|P_n\| = 1$, and therefore, using Remark 3.25 and (11.33) we can estimate

$$\|A_n^{-1} P_n A\| \le \frac{1}{c} \, \|A\|,$$

whence (11.32) follows. □

The error estimate of Theorem 11.17 is usually referred to as *Céa's lemma*, since it was first obtained by Céa in 1964. It indicates that the error in the Galerkin method is determined by how well the exact solution can be approximated by elements of the subspace X_n.

By Corollary 11.16 the Galerkin method immediately carries over to the solution of the sesquilinear equation (11.27) and consists in finding $u_n \in X_n$ such that

$$S(v, u_n) = F(v) \tag{11.34}$$

for all $v \in X_n$.

The practical solution of the Galerkin equations (11.30) reduces to the solution of a system of linear equations. If $w_1, \ldots, w_n$ is a basis for X_n (without loss of generality we assume the dimension of X_n to be n), then for

$$u_n = \sum_{k=1}^{n} \alpha_k w_k$$

the Galerkin equations (11.30) are equivalent to the system of linear equations

$$\sum_{k=1}^{n} \bar{\alpha}_k (w_j, A w_k) = (w_j, f), \quad j = 1, \ldots, n. \tag{11.35}$$

From this formulation it becomes obvious that the Galerkin method is only a *semidiscrete method*, since setting up the linear system requires the evaluation of scalar products and of the operator A applied to the basis elements. For a *fully discrete method* these computations, in general, need further approximations of integrals for the scalar products and of differential or integral operators. This also requires that the error analysis be amended accordingly, since the error estimate of Theorem 11.17 covers only the semidiscrete case.

Having outlined the basic ideas of the Galerkin method and its error analysis within a few paragraphs, we want to point out clearly that the power and the art of the application of the Galerkin method for the approximate solution of differential and integral equations begins with the proper choice of the approximating subspace X_n and the appropriate basis $w_1, \dots, w_n$ therein, corresponding to the operator A under consideration. However, it is beyond our goal to enter into this important topic in any detail aside from the short discussion in Section 11.5.

11.4 Weak Solutions

We return to the boundary value problem, and instead of (11.7)–(11.8) we consider the slightly more general so-called *Sturm–Liouville problem*

$$-(pu')' + qu = r \quad \text{in } [a, b] \tag{11.36}$$

with homogeneous boundary conditions

$$u(a) = u(b) = 0. \tag{11.37}$$

Here we assume that $p \in C^1[a, b]$ and $q, r \in C[a, b]$ such that $p(x) > 0$ and $q(x) \geq 0$ for all $x \in [a, b]$. Multiplying the differential equation by v and performing a partial integration, it follows that each solution u to (11.36)–(11.37) satisfies

$$S(v, u) = F(v) \tag{11.38}$$

for all $v \in C^1[a, b]$ with $v(a) = v(b) = 0$, where we have set

$$S(u, v) := \int_a^b (pu'v' + quv)\, dx \tag{11.39}$$

and

$$F(v) := \int_a^b rv\, dx. \tag{11.40}$$

Conversely, if $u \in C^2[a, b]$ satisfies (11.38), by partial integration we obtain that

$$\int_a^b [(pu')' - qu + r]v\, dx = 0 \tag{11.41}$$

for all $v \in C^1[a,b]$ with $v(a) = v(b) = 0$. Now we set $f := (pu')' - qu + r$ and assume that $f(x_0) \neq 0$ for some x_0 in (a,b), say $f(x_0) > 0$. Since f is continuous, there exists an interval $U \subset (a,b)$ such that f is positive on U. Now we choose a nonnegative function $v \neq 0$ from $C^1[a,b]$ which vanishes outside U. For this function v the integral in (11.41) must be positive. This is a contradiction, and therefore f must vanish identically; i.e., u satisfies the differential equation (11.36). Therefore, (11.38) provides an equivalent reformulation of the boundary value problem.

From Example 3.38 we recall that the space of continuous functions is not complete with respect to the L_2 scalar product. However, if we wish to apply the analysis of the previous section and, in particular, Corollary 11.16, then we need a Hilbert space. For this, we introduce the *Sobolev space* $H^1[a,b]$ based on the concept of weak derivatives. By $L^2[a,b]$ we denote the space of measurable real-valued functions defined on the interval $[a,b]$ that are square-integrable in the sense of Lebesgue. We shall make use of the fact that $L^2[a,b]$ is a Hilbert space with respect to the L_2 scalar product. (More precisely, $L^2[a,b]$ is the linear space of equivalence classes of functions coinciding almost everywhere.) Note that the space $C[a,b]$ of continuous functions is dense in $L^2[a,b]$ (see [5, 51, 59]).

Definition 11.18 *A function $u \in L^2[a,b]$ is said to have a* weak derivative *$u' \in L^2[a,b]$ if*

$$\int_a^b uv'\,dx = -\int_a^b u'v\,dx \tag{11.42}$$

for all $v \in C^1[a,b]$ with $v(a) = v(b) = 0$.

By partial integration it follows that (11.42) is satisfied for functions $u \in C^1[a,b]$. Hence, weak differentiability generalizes classical differentiability.

From the denseness of $\{v \in C^1[a,b] : v(a) = v(b) = 0\}$ in $L^2[a,b]$, or from the Fourier series for the odd extension of u, it can be seen that the weak derivative, if it exists, is unique (see Problem 11.17). From the denseness of $C[a,b]$ in $L^2[a,b]$, or from the Fourier series for the even extension of u, it follows that each function with vanishing weak derivative must be constant almost everywhere (see Problem 11.17). The latter, in particular, implies

$$u(x) = \int_a^x u'(\xi)\,d\xi + c \tag{11.43}$$

for almost all $x \in [a,b]$ and some constant c, since by Fubini's theorem

$$\begin{aligned}\int_a^b \left(\int_a^x u'(\xi)\,d\xi\right) v'(x)\,dx &= \int_a^b u'(\xi)\left(\int_\xi^b v'(x)\,dx\right) d\xi\\ &= -\int_a^b u'(\xi)v(\xi)\,d\xi\end{aligned}$$

for all $v \in C^1[a,b]$ with $v(a) = v(b) = 0$. Hence both sides of (11.43) have the same weak derivative.

Theorem 11.19 *The linear space*

$$H^1[a,b] := \{u \in L^2[a,b] : u' \in L^2[a,b]\}$$

endowed with the scalar product

$$(u,v)_{H^1} := \int_a^b (uv + u'v')\,dx \tag{11.44}$$

is a Hilbert space.

Proof. It is readily checked that $H^1[a,b]$ is a linear space and that (11.44) defines a scalar product. Let (u_n) denote an H^1 Cauchy sequence. Then (u_n) and (u'_n) are both L^2 Cauchy sequences. From the completeness of $L^2[a,b]$ we obtain the existence of $u \in L^2[a,b]$ and $w \in L^2[a,b]$ such that $\|u_n - u\|_2 \to 0$ and $\|u'_n - w\|_2 \to 0$ as $n \to \infty$. Then for all $v \in C^1[a,b]$ with $v(a) = v(b) = 0$ we can estimate

$$\int_a^b (uv' + wv)\,dx = \int_a^b \{(u - u_n)v' + (w - u'_n)v\}\,dx$$

$$\leq \|u - u_n\|_{L^2}\|v'\|_{L^2} + \|w - u'_n\|_{L^2}\|v\|_{L^2} \to 0, \quad n \to \infty.$$

Therefore, $u \in H^1[a,b]$ with $u' = w$, and $\|u - u_n\|_{H^1} \to 0$, $n \to \infty$, which completes the proof. □

Theorem 11.20 *$C^1[a,b]$ is dense in $H^1[a,b]$.*

Proof. Since $C[a,b]$ is dense in $L^2[a,b]$, for each $u \in H^1[a,b]$ and $\varepsilon > 0$ there exists $w \in C[a,b]$ such that $\|u' - w\|_2 < \varepsilon$. Then we define $v \in C^1[a,b]$ by

$$v(x) := u(a) + \int_a^x w(\xi)\,d\xi,$$

and using (11.43), we have

$$u(x) - v(x) = \int_a^x \{u'(\xi) - w(\xi)\}\,d\xi.$$

By the Cauchy–Schwarz inequality this implies $\|u - v\|_2 < (b - a)\varepsilon$, and the proof is complete. □

Theorem 11.21 *$H^1[a,b]$ is contained in $C[a,b]$.*

Proof. From (11.43) we have

$$u(x) - u(y) = \int_y^x u'(\xi)\, d\xi, \tag{11.45}$$

whence by the Cauchy–Schwarz inequality,

$$|u(x) - u(y)| \leq |x - y|^{1/2} \|u'\|_2$$

follows for all $x, y \in [a, b]$. Therefore, every function $u \in H^1[a, b]$ belongs to $C[a, b]$, or more precisely, it coincides almost everywhere with a continuous function. □

By Theorem 11.21 we may consider $H^1[a, b]$ as a subspace of $C[a, b]$. Choose $y \in [a, b]$ such that $|u(y)| = \min_{a \leq x \leq b} |u(x)|$. Then from

$$(b - a) \min_{a \leq x \leq b} |u(x)|^2 \leq \int_a^b |u(x)|^2 dx$$

and (11.45), by the Cauchy–Schwarz inequality we find that

$$\|u\|_\infty \leq C\|u\|_{H^1}$$

for some constant C. The latter inequality means that the H^1 norm is stronger than the maximum norm (in one space dimension!).

Theorem 11.22 *The space*

$$H_0^1[a, b] := \{u \in H^1[a, b] : u(a) = u(b) = 0\}$$

is a complete subspace of $H^1[a, b]$.

Proof. Since the H^1 norm is stronger than the maximum norm, each H^1 convergent sequence of elements of $H_0^1[a, b]$ has its limit in $H_0^1[a, b]$. Therefore $H_0^1[a, b]$ is a closed subspace of $H^1[a, b]$, and the statement follows from Remark 3.40. □

Definition 11.23 *A function $u \in H_0^1[a, b]$ is called a* weak solution *to the boundary value problem (11.36)–(11.37) if (11.38) is satisfied for all $v \in H_0^1[a, b]$.*

Theorem 11.24 *Assume that $p > 0$ and $q \geq 0$. Then there exists a unique weak solution to the boundary value problem (11.36)–(11.37).*

Proof. The sesquilinear function $S : H_0^1[a, b] \times H_0^1[a, b]$ is bounded, since

$$|S(u, v)| \leq \max\left\{\|p\|_\infty, \|q\|_\infty\right\} \|u\|_{H^1} \|v\|_{H^1}$$

by the Cauchy–Schwarz inequality. For $u \in H_0^1[a,b]$, from (11.45) and the Cauchy–Schwarz inequality we obtain that

$$||u||_{L^2}^2 = \int_a^b \left| \int_a^x u'(\xi)\, d\xi \right|^2 dx \le (b-a)^2 ||u'||_{L^2}^2.$$

Hence we can estimate

$$S(u,u) \ge \min_{a \le x \le b} p(x) \int_a^b |u'|^2\, dx \ge c||u||_{H^1}^2$$

for all $u \in H_0^1[a,b]$ and some positive constant c; i.e., S is strictly coercive. Finally, by the Cauchy–Schwarz inequality we have

$$|F(v)| \le ||r||_{L^2} ||v||_{L^2} \le ||r||_{L^2} ||v||_{H^1};$$

i.e., the linear function $F : H_0^1[a,b] \to \mathbb{R}$ is bounded. Now the statement of the theorem follows from Corollary 11.16. □

We note that from (11.28) and the previous inequality it follows that

$$||u||_{H^1} \le \frac{1}{c}\, ||r||_{L^2} \tag{11.46}$$

for the weak solution u to the boundary value problem (11.36)–(11.37).

Theorem 11.25 *Each weak solution to the boundary value problem (11.36)–(11.37) is also a classical solution; i.e., it is twice continuously differentiable.*

Proof. Define

$$f(x) := \int_a^x [q(\xi)u(\xi) - r(\xi)]\, d\xi, \quad x \in [a,b].$$

Then $f \in C^1[a,b]$. From (11.38), by partial integration we obtain

$$\int_a^b [pu' - f]v'\, dx = 0$$

for all $v \in H_0^1[a,b]$. Now we set

$$c := \frac{1}{b-a} \int_a^b [pu' - f]\, d\xi$$

and

$$v_0(x) := \int_a^x [p(\xi)u'(\xi) - f(\xi) - c]\, d\xi, \quad x \in [a,b].$$

Then $v_0 \in H_0^1[a,b]$ and

$$\int_a^b [pu' - f - c]^2 \, dx = \int_a^b [pu' - f - c] v_0' \, dx$$

$$= \int_a^b [pu' - f] v_0' \, dx - c \int_a^b v_0' \, dx = 0.$$

Hence

$$pu' = f + c,$$

and since f and p are in $C^1[a,b]$ with $p(x) > 0$ for all $x \in [a,b]$, we can conclude that $u' \in C^1[a,b]$ and

$$(pu')' = f' = qu - r.$$

This completes the proof. □

Using the differential equation (11.36), from (11.46) we conclude that there exists a constant $C > 0$, independent of r, such that

$$\|u''\|_{L^2} \le C \|r\|_{L^2}, \tag{11.47}$$

which we note for later use.

As compared to Theorem 11.4 we have not obtained any major extension of the existence result. However, as pointed out already in the introduction, we view this section as a model case for the more complicated situation of partial differential equations. By partial integration it can be seen that the boundary value problem (11.17)–(11.18) for the Laplace operator is equivalent to finding a function $u \in C^2(D)$ satisfying $u = 0$ on ∂D and

$$\int_D ([\operatorname{grad} u]^T \operatorname{grad} v + quv) dx = \int_D fv \, dx \tag{11.48}$$

for all $v \in C^1(D)$ with $v = 0$ on ∂D. The analysis of weak solutions of (11.48) follows the same pattern as for the ordinary differential equation (11.36). However, the details are more heavily involved. In particular, since for the multidimensional case the Sobolev space $H^1(D)$ no longer is a subspace of the continuous functions, the formulation of the boundary condition, i.e., the definition of the subspace $H_0^1(D)$, has to be modified, and establishing that weak solutions are also classical solutions is more complicated. For a comprehensive study of weak solutions to boundary value problems for elliptic partial differential equations we refer to [24, 60].

11.5 The Finite Element Method

The *finite element* method for the boundary value problem (11.36)–(11.37) consists in the application of the Galerkin method (11.34) to the weak

formulation (11.38) by using spline spaces as approximating subspaces. Then, for appropriate basis functions, the matrix $S(w_j, w_k)$ will be sparse; i.e., most of the matrix entries will be zero. Polynomials as approximating subspaces are not suitable, since analogously to Example 5.1, they lead to ill-conditioned linear systems with full matrices.

We consider the case of linear splines. For the equidistant grid

$$x_j := a + jh, \quad j = 0, \ldots, n+1,$$

with step size $h = (b-a)/(n+1)$ and $n \in \mathbb{N}$ we choose for X_n the space of continuous piecewise linear functions; i.e., X_n consists of the functions $u \in C[a,b]$ that satisfy $u(a) = u(b) = 0$ and coincide on each subinterval $[x_{j-1}, x_j]$ with a polynomial in P_1 for $j = 1, \ldots, n$. The functions in this spline space belong to $H_0^1[a,b]$ with piecewise constant weak derivatives. As basis elements in X_n we take the so-called *hat functions*

$$w_k(x) := \begin{cases} \dfrac{1}{h}(x - x_{k-1}), & x \in [x_{k-1}, x_k], \\[2ex] \dfrac{1}{h}(x_{k+1} - x), & x \in [x_k, x_{k+1}], \\[2ex] 0, & x \notin [x_{k-1}, x_{k+1}]. \end{cases}$$

Each $u \in X_n$ can be represented in the form

$$u = \sum_{k=1}^{n} \alpha_k w_k,$$

where $\alpha_k = u(x_k)$, $k = 1, \ldots, n$. Obviously, we have

$$S(w_j, w_k) = \int_a^b \{p w_j' w_k' + q w_j w_k\}\, dx = 0$$

if $(x_{j-1}, x_{j+1}) \cap (x_{k-1}, x_{k+1}) = \emptyset$, i.e., if $|j-k| > 2$. Therefore, the matrix $S(w_j, w_k)$ is tridiagonal. We compute the matrix elements

$$\begin{aligned} S(w_j, w_j) = {} & \frac{1}{h^2} \int_{x_{j-1}}^{x_{j+1}} p(x)\, dx \\ & + \frac{1}{h^2} \left\{ \int_{x_{j-1}}^{x_j} q(x)(x - x_{j-1})^2 dx + \int_{x_j}^{x_{j+1}} q(x)(x_{j+1} - x)^2 dx \right\} \end{aligned}$$

and

$$\begin{aligned} S(w_j, w_{j+1}) &= S(w_{j+1}, w_j) \\ &= \frac{-1}{h^2} \int_{x_j}^{x_{j+1}} p(x)\, dx + \frac{1}{h^2} \int_{x_j}^{x_{j+1}} q(x)(x_{j+1} - x)(x - x_j)\, dx, \end{aligned}$$

and the right-hand sides

$$F(w_j) = \frac{1}{h}\left\{\int_{x_{j-1}}^{x_j} r(x)(x - x_{j-1})\,dx + \int_{x_j}^{x_{j+1}} r(x)(x_{j+1} - x)\,dx\right\}.$$

These equations illustrate two general features of the finite element methods. Firstly, it is characteristic for the finite element method that the coefficients are computed by the same formula for each subinterval, i.e., for each of the finite elements into which the total interval is subdivided.

Secondly, as already mentioned earlier, the Galerkin method is only semidiscrete. In order to make it fully discrete, a numerical quadrature has to be applied. If we remain within our framework of approximations and approximate p, q, and r by linear splines, we obtain

$$S(w_j, w_j) \approx \frac{1}{2h}\,(p_{j-1} + 2p_j + p_{j+1}) + \frac{h}{12}\,(q_{j-1} + 6q_j + q_{j+1})$$

and

$$S(w_j, w_{j+1}) \approx \frac{-1}{2h}\,(p_j + p_{j+1}) + \frac{h}{12}\,(q_j + q_{j+1})$$

for the matrix elements, and

$$F(w_j) \approx \frac{h}{6}(r_{j-1} + 4r_j + r_{j+1})$$

for the right-hand sides. Here, as above, we have set $p_j = p(x_j)$, $q_j = q(x_j)$, and $r_j = r(x_j)$. Similar to the linear system (11.10)–(11.11) for the finite difference method, the tridiagonal linear system is irreducible and weakly row-diagonally dominant. It also is accessible to convergence acceleration of the Jacobi iterations by relaxation and multigrid methods.

In order to derive an error estimate for the semidiscrete version of the finite element method with linear splines from Theorem 11.17, we need an estimate for the interpolation error for linear splines with respect to the H^1 norm (see also Theorem 8.33).

Lemma 11.26 *Let $f[a,b] \in C^2[a,b]$. Then the remainder $R_1 f := f - L_1 f$ for the linear interpolation at the two endpoints a and b can be estimated by*

$$\begin{aligned} \|R_1 f\|_{L^2} &\le (b-a)^2 \|f''\|_{L^2}, \\ \|(R_1 f)'\|_{L^2} &\le (b-a)\,\|f''\|_{L^2}. \end{aligned} \tag{11.49}$$

Proof. For each function $g \in C^1[a,b]$ satisfying $g(a) = 0$, from

$$g(x) = \int_a^x g'(\xi)\,d\xi,$$

by using the Cauchy–Schwarz inequality we obtain

$$|g(x)|^2 \le (b-a)\|g'\|_{L^2}^2, \quad x \in [a,b].$$

From this, by integration we derive the *Friedrich inequality*

$$\|g\|_{L^2} \le (b-a)\|g'\|_{L^2} \tag{11.50}$$

for functions $g \in C^1[a,b]$ with $g(a) = 0$ (or $g(b) = 0$). Using the interpolation property $(R_1 f)(a) = (R_1 f)(b) = 0$, by partial integration we obtain

$$\int_a^b [f' - (L_1 f)']^2 dx = \int_a^b f''(L_1 f - f)\, dx.$$

From this, again applying the Cauchy–Schwarz inequality, we have

$$\|(R_1 f)'\|_{L^2}^2 \le \|f''\|_{L^2}\, \|R_1 f\|_{L^2},$$

whence (11.49) follows with the aid of Friedrich's inequality (11.50) for $g = R_1 f$. □

Theorem 11.27 *The error in the finite element approximation by linear splines for the boundary value problem (11.36)–(11.37) can be estimated by*

$$\|u_n - u\|_{H^1} \le C\|u''\|_{L^2} h \tag{11.51}$$

for some positive constant C.

Proof. By summing up the inequalities (11.49), applied to each of the subintervals of length h, for the interpolating linear spline $w_n \in X_n$ with $w_n(x_j) = u(x_j)$ for $j = 0, \dots, n$ we find that

$$\|w_n' - u'\|_{L^2} \le \|u''\|_{L^2} h$$

and

$$\|w_n - u\|_{L^2} \le \|u''\|_{L^2} h^2,$$

whence

$$\inf_{v \in X_n} \|v - u\|_{H^1} \le \|w_n - u\|_{H^1} \le (1 + b - a)\|u''\|_{L^2} h$$

follows. Now (11.51) is a consequence of the error estimate for the Galerkin method of Theorem 11.17. □

By the following trick, which was independently developed by Aubin (1967) and Nitsche (1968), we can improve the error estimate in the L_2 norm to the order $O(h^2)$ that we expect for approximations using linear splines.

Theorem 11.28 *The error in the finite element approximation by linear splines for the boundary value problem (11.36)–(11.37) can be estimated by*

$$\|u_n - u\|_{L^2} \le C\|u''\|_{L^2} h^2$$

with some positive constant C.

Proof. Denote by z_n the weak solution to the boundary value problem with the right-hand side $u - u_n$; i.e.,

$$S(v, z_n) = (v, u - u_n)_{L^2}$$

for all $v \in H_0^1[a, b]$. In particular, inserting $v = u - u_n$, it follows that

$$S(u - u_n, z_n) = \|u - u_n\|_{L^2}^2. \tag{11.52}$$

Since $S(v, u) = F(v)$ and $S(v, u_n) = F(v)$ for all $v \in X_n$, using the symmetry of S we have

$$S(u - u_n, v) = 0$$

for all $v \in X_n$. Inserting the Galerkin approximation to z_n, which we denote by $\tilde{z}_n$, into the last equation and subtracting from (11.52), we obtain

$$\|u - u_n\|_{L^2}^2 = S(u - u_n, z_n - \tilde{z}_n). \tag{11.53}$$

Since S is bounded, from (11.53) and (11.51), applied to $u - u_n$ and $z_n - \tilde{z}_n$, we can conclude that

$$\|u - u_n\|_{L^2}^2 \le C_1 \|u''\|_{L^2} \|z_n''\|_{L^2} h^2$$

for some constant C_1. However, from (11.47) we also have that

$$\|z_n''\|_{L^2} \le C_2 \|u - u_n\|_{L^2}$$

for some constant C_2. Now the assertion of the theorem follows from the last two inequalities. □

We refrain from describing both the extension of this analysis to higher-order splines such as cubic splines (see Problem 11.19) and the extension to partial differential equations. For the latter we refer to [4, 11].

Problems

11.1 Consider multiple shooting for the boundary value problem

$$u'' + u = 0, \quad u(a) = u(b) = 0,$$

with n equidistant subintervals. Show that the corresponding linear system (11.6) is uniquely solvable, provided that $(b - a)/\pi \notin \mathbb{N}$.

11.2 Write a computer program for multiple shooting using the Newton method and the Runge–Kutta method and test it for various examples.

11.3 Show that the boundary value problem for the differential equation

$$u'' = f(x, u, u'), \quad a \le x \le b,$$

with inhomogeneous boundary conditions $u(a) = \alpha$ and $u(b) = \beta$ can be equivalently transformed into a boundary value problem with homogeneous boundary condition.

11.4 Show that the general solution of the linear differential equation (11.7) is given by (11.9).

11.5 For $p \in C^1[a, b]$ and $q \in C[a, b]$ show that the boundary value problem

$$u'' + pu' + qu = r \quad \text{in } [a, b], \quad u(a) = u(b) = 0,$$

is solvable for each right-hand side $r \in C[a, b]$ if and only if the boundary value problem

$$u'' + pu' + qu = 0 \quad \text{in } [a, b], \quad u(a) = u(b) = 0,$$

admits only the trivial solution $u = 0$.

11.6 Find the solution of the boundary value problem

$$u''(x) + u(x) = e^x, \quad u(0) = u(1) = 0.$$

11.7 Write a computer program for the finite difference method (11.10)–(11.11) and test it for various examples.

11.8 Find the explicit solution for the finite difference approximation (11.10)–(11.11) for the boundary value problem

$$u'' - u = -2 \quad \text{in } [0, 1], \quad u(0) = u(1) = 0,$$

and verify the convergence result of Theorem 11.8.

11.9 Show that the error in the finite difference approximation (11.15) is of order $O(h^2)$ and that the error in the approximation (11.16) is of order $O(h^4)$.

11.10 Prove the estimate $\|u_0\|_\infty \le 1/8$ for the solution to the boundary value problem (11.21).

11.11 In the space $C[a, b]$ with scalar product

$$(u, v) := \int_a^b u(x)\overline{v(x)}\, dx,$$

define a functional $F : C[a, b] \to \mathbb{C}$ by

$$F(u) := \int_a^b u(x)\, dx.$$

Show that F is linear and bounded. Is there an $f \in C[a, b]$ such that $F(u) = (u, f)$ for all $u \in C[a, b]$? Does your answer agree with the Riesz Theorem 11.11?

11.12 In the pre-Hilbert space of Problem 11.11, for a fixed $x \in [a,b]$ consider the point evaluation functional $F : C[a,b] \to \mathbb{C}$ defined by

$$F(u) := u(x).$$

Is F linear and bounded?

11.13 Let X and Y be Hilbert spaces and let $A : X \to Y$ be a bounded linear operator. Show that there exists a uniquely determined bounded linear operator $A^* : Y \to X$ such that

$$(Au, v) = (u, A^*v)$$

for all $u \in X$ and $v \in Y$. The operator A^* is called the *adjoint operator* of A. Show that $\|A\| = \|A^*\|$.

11.14 Let $A : X \to X$ be a bounded, self-adjoint, and positive operator in a Hilbert space X; i.e., $(Au, v) = (u, Av)$ for all $u, v \in X$ and $(Au, u) > 0$ for all $u \neq 0$. Choose $w_0 \in X$ and define $w_j = Aw_{j-1}$ for $j = 1, \dots, n-1$. Show that the Galerkin equations for $Au = f$ with respect to the subspaces $X_n = \operatorname{span}\{w_0, \dots, w_{n-1}\}$ are uniquely solvable for each $n \in \mathbb{N}$. Moreover, if f is in the closure of $\operatorname{span}\{A^j w_0 : j = 0, 1, \dots\}$, then the Galerkin approximation u_n converges to the solution of $Au = f$.

Show that in the special case $w_0 = f$ the approximations u_n can be computed iteratively by the formulae $u_0 = 0$, $p_0 = f$, and

$$\begin{aligned}
u_{n+1} &= u_n - \alpha_n p_n, \\
p_n &= r_n + \beta_{n-1} p_{n-1}, \\
r_n &= r_{n-1} - \alpha_{n-1} A p_{n-1}, \\
\alpha_{n-1} &= (r_{n-1}, p_{n-1})/(Ap_{n-1}, p_{n-1}), \\
\beta_{n-1} &= -(r_n, Ap_{n-1})/(Ap_{n-1}, p_{n-1}).
\end{aligned}$$

Here r_n is the residual $r_n = Au_n - f$. This is the *conjugate gradient method* of Hestenes and Stiefel.

11.15 Let $A : X \to X$ be a bounded, self-adjoint, and positive operator in a Hilbert space X; i.e., $(Au, v) = (u, Av)$ for all $u, v \in X$ and $(Au, u) > 0$ for all $u \neq 0$. Show that solving the equation $Au = f$ is equivalent to minimizing the so-called *energy functional*

$$E(v) := (v, Av) - 2\,\mathrm{Re}(v, f)$$

on X. Show that the Galerkin approximation with respect to a subspace X_n is equivalent to minimizing E on X_n. This method is known as the *Rayleigh–Ritz method.*

11.16 Show that under the assumptions of Problem 11.15 for the Galerkin equations the SOR method of Section 4.2 converges for $0 < \omega < 2$.

11.17 Show that the weak derivative, if it exists, is unique and that each function with vanishing weak derivative must be constant almost everywhere.

11.18 Write a computer program for the finite element method with linear splines and test it for various examples. Compare the numerical results with those for the finite difference method.

11.19 Let $B_{-1}, B_0, B_1, \ldots, B_n, B_{n+1}, B_{n+2}$ denote the cubic B-splines for the equidistant grid $x_j := a + jh$, $j = 0, \ldots, n+1$, with step size $h = (b-a)/n$. Show that

$$u_0 := B_0 - 4B_{-1}, \ u_1 := B_1 - B_{-1},$$

$$u_2 := B_2, \ldots, u_{n-1} := B_{n-1},$$

$$u_n := B_n - B_{n+2}, \ u_{n+1} := B_n - 4B_{n+2}$$

is a basis for $S_3^n \cap H_0^1[a,b]$, i.e., for the space of cubic splines vanishing at the endpoints.

Using this basis, set up the Galerkin equations for the Sturm–Liouville problem analogous to the case of linear splines treated in Section 11.5.

11.20 Formulate and prove analogues of Theorems 11.27 and 11.28 for the finite element approximation using cubic splines as in Problem 11.19.

12

Integral Equations

The topic of the last chapter of this book is linear integral equations, of which

$$\int_a^b K(x,y)\varphi(y)\,dy = f(x), \quad x \in [a,b],$$

and

$$\varphi(x) - \int_a^b K(x,y)\varphi(y)\,dy = f(x), \quad x \in [a,b],$$

are typical examples. In these equations the function φ is the unknown, and the so-called *kernel* K and the right-hand side f are given functions. The above equations are called *Fredholm integral equations* of the *first* and *second kind*, respectively. Since both the theory and the numerical approximations for integral equations of the first kind are far more complicated than for integral equations of the second kind, we will confine our presentation to the latter case.

Integral equations provide an important tool for solving boundary value problems for both ordinary and partial differential equations (see Problem 12.1 and [39]). Their historical development is closely related to the solution of boundary value problems in potential theory in the last decades of the nineteenth century. Progress in the theory of integral equations also had a great impact on the development of functional analysis.

Omitting the proofs, we will present the main results of the Riesz theory for compact operators as the foundation of the existence theory for integral equations of the second kind. Then we will develop the fundamental ideas of the Nyström method and the collocation method as the two most im-

portant approaches for the numerical solution of these integral equations. This is done in a general framework of operator equations and their approximate solution, which makes the analysis more widely applicable. For a comprehensive study of both the theory and the numerical solution of linear integral equations we refer to [39].

12.1 The Riesz Theory

This section is devoted to a summary of some of the basic facts of the theory of Fredholm integral equations of the second kind. The integral equations formulated above carry the name of Fredholm, since in 1902 Fredholm established an existence theory for integral equations of the second kind with continuous kernels, which is now known as the Fredholm alternative. For the purpose of this introduction to the numerical solution of integral equations it suffices to consider only the first and most important part of this alternative, which states that the inhomogeneous equation

$$\varphi(x) - \int_a^b K(x,y)\varphi(y)\,dy = f(x), \quad x \in [a,b], \tag{12.1}$$

with continuous kernel K has a unique solution $\varphi \in C[a,b]$ for each right-hand side $f \in C[a,b]$ if and only if the homogeneous integral equation

$$\varphi(x) - \int_a^b K(x,y)\varphi(y)\,dy = 0, \quad x \in [a,b], \tag{12.2}$$

has only the trivial solution. The importance of this result originates from the fact that it reduces the difficult problem of establishing existence of a solution to the inhomogeneous integral equation to the simpler problem of showing that the homogeneous integral equation allows only the trivial solution $\varphi = 0$, and it extends the corresponding statement for systems of linear equations to the case of integral equations. Actually, Fredholm derived his results by interpreting integral equations as a limiting case of linear systems by considering the integral as a limit of Riemann sums and passing to the limit in Cramer's rule for the solution of linear systems. For the solution of integral equations with continuous kernels, Fredholm's approach is still the most elegant and shortest. However, since it is restricted to the case of continuous kernels, it is more convenient to consider the above equations as a special case of operator equations of the second kind with a compact operator, as presented by Riesz in 1918.

Definition 12.1 *A linear operator $A : X \to Y$ from a normed space X into a normed space Y is called* compact *if for each bounded sequence (φ_n) in X the sequence $(A\varphi_n)$ contains a convergent subsequence in Y, i.e., if each sequence from the set $\{A\varphi : \varphi \in X,\ \|\varphi\| \leq 1\}$ contains a convergent subsequence.*

Without developing the concept of compactness in normed spaces in any detail, we note that this definition is equivalent to requiring that the set $\{A\varphi : \varphi \in X, \|\varphi\| \leq 1\}$ be *relatively sequentially compact.*

Compact operators are bounded, linear combinations of compact operators are compact, and products of two bounded operators are compact if one of them is compact (see Problem 12.2). From the Bolzano–Weierstrass theorem it can be seen that bounded operators $A : X \to X$ with finite-dimensional range $A(X) := \{A\varphi : \varphi \in X\}$ are compact. Furthermore, the identity operator $I : X \to X$, defined by $I : \varphi \mapsto \varphi$ for all $\varphi \in X$, is compact if and only if the space X is finite-dimensional. This actually justifies the distinction between the equations $A\varphi = f$ and $\varphi - A\varphi = f$ as equations of the first and second kind, since A and $I - A$ have different properties in infinite-dimensional spaces if A is compact. A proof of these facts and of the following important theorem can be found in most introductory books on functional analysis, for example in [39].

The fundamental result of the Riesz theory is described by the following theorem, which extends Fredholm's result on the equivalence of injectivity and surjectivity to the case of operator equations of the second kind with a compact operator.

Theorem 12.2 *Let $A : X \to X$ be a compact operator in a normed space X. Then $I - A$ is surjective if and only if it is injective. If the inverse operator $(I - A)^{-1} : X \to X$ exists, it is bounded.*

In order to verify that Fredholm's existence analysis for integral equations with continuous kernels $K : [a, b] \times [a, b] \to \mathbb{R}$ can be viewed as a special case of Theorem 12.2, we have to establish that the linear integral operator $A : C[a, b] \to C[a, b]$, defined by

$$(A\varphi)(x) := \int_a^b K(x, y)\varphi(y)\, dy, \quad x \in [a, b], \tag{12.3}$$

is compact. For this we need the following theorem due to Arzelà–Ascoli, which again is proven in most introductions to functional analysis.

Theorem 12.3 (Arzelà–Ascoli) *Each sequence from a subset $U \subset C[a, b]$ contains a uniformly convergent subsequence; i.e., U is relatively sequentially compact, if and only if it is bounded and* equicontinuous, *i.e., if there exists a constant C such that*

$$|\varphi(x)| \leq C$$

for all $x \in [a, b]$ and all $\varphi \in U$, and for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|\varphi(x) - \varphi(y)| < \varepsilon$$

for all $x, y \in [a, b]$ with $|x - y| < \delta$ and all $\varphi \in U$.

Theorem 12.4 *The integral operator (12.3) with continuous kernel is a compact operator on* $C[a,b]$.

Proof. For all $\varphi \in C[a,b]$ with $\|\varphi\|_\infty \le 1$ and all $x \in [a,b]$, we have that

$$|(A\varphi)(x)| \le (b-a) \max_{x,y\in[a,b]} |K(x,y)|;$$

i.e., the set $U := \{A\varphi : \varphi \in C[a,b], \|\varphi\|_\infty \le 1\} \subset C[a,b]$ is bounded. Since K is uniformly continuous on the square $[a,b] \times [a,b]$, for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$|K(x,z) - K(y,z)| < \frac{\varepsilon}{b-a}$$

for all $x, y, z \in [a,b]$ with $|x-y| < \delta$. Then

$$|(A\varphi)(x) - (A\varphi)(y)| = \left| \int_a^b [K(x,z) - K(y,z)]\varphi(z)\,dz \right| < \varepsilon$$

for all $x, y \in [a,b]$ with $|x-y| < \delta$ and all $\varphi \in C[a,b]$ with $\|\varphi\|_\infty \le 1$; i.e., U is equicontinuous. Hence A is compact by the Arzelà–Ascoli Theorem 12.3. □

In our analysis we also will need an explicit expression for the norm of the integral operator A.

Theorem 12.5 *The norm of the integral operator* $A : C[a,b] \to C[a,b]$ *with continuous kernel* K *is given by*

$$\|A\|_\infty = \max_{a\le x\le b} \int_a^b |K(x,y)|\,dy. \tag{12.4}$$

Proof. For each $\varphi \in C[a,b]$ with $\|\varphi\|_\infty \le 1$ we have

$$|(A\varphi)(x)| \le \int_a^b |K(x,y)|\,dy, \quad x \in [a,b],$$

and thus

$$\|A\|_\infty = \sup_{\|\varphi\|_\infty \le 1} \|A\varphi\|_\infty \le \max_{a\le x\le b} \int_a^b |K(x,y)|\,dy.$$

Since K is continuous, there exists $x_0 \in [a,b]$ such that

$$\int_a^b |K(x_0,y)|\,dy = \max_{a\le x\le b} \int_a^b |K(x,y)|\,dy.$$

For $\varepsilon > 0$ choose $\psi \in C[a,b]$ by setting

$$\psi(y) := \frac{K(x_0,y)}{|K(x_0,y)| + \varepsilon}, \quad y \in [a,b].$$

Then $||\psi||_\infty \leq 1$ and

$$||A\psi||_\infty \geq |(A\psi)(x_0)| = \int_a^b \frac{[K(x_0,y)]^2}{|K(x_0,y)| + \varepsilon}\, dy \geq \int_a^b \frac{[K(x_0,y)]^2 - \varepsilon^2}{|K(x_0,y)| + \varepsilon}\, dy$$

$$= \int_a^b |K(x_0,y)|\, dy - \varepsilon(b-a).$$

Hence

$$||A||_\infty = \sup_{||\varphi||_\infty \leq 1} ||A\varphi||_\infty \geq ||A\psi||_\infty \geq \int_a^b |K(x_0,y)|\, dy - \varepsilon(b-a),$$

and since this holds for all $\varepsilon > 0$, we have

$$||A||_\infty \geq \int_a^b |K(x_0,y)|\, dy = \max_{a \leq x \leq b} \int_a^b |K(x,y)|\, dy.$$

This concludes the proof. □

It also can be shown that the integral operator remains compact if the kernel K is merely weakly singular (see [39]). A kernel K is said to be *weakly singular* if it is defined and continuous for all $x, y \in [a,b]$, $x \neq y$, and there exist positive constants M and $\alpha \in (0,1]$ such that

$$|K(x,y)| \leq M|x-y|^{\alpha-1}$$

for all $x, y \in [a,b]$, $x \neq y$.

12.2 Operator Approximations

The fundamental concept for approximately solving an operator equation

$$\varphi - A\varphi = f$$

of the second kind is to replace it by an equation

$$\varphi_n - A_n\varphi_n = f_n$$

with approximating sequences $A_n \to A$ and $f_n \to f$ as $n \to \infty$. For computational purposes, the approximating equations will be chosen such that

they can be reduced to solving a system of linear equations. In this section we will provide a convergence and error analysis for such approximation schemes. In particular, we will derive convergence results and error estimates for the cases where we have either norm or pointwise convergence of the sequence $A_n \to A$, $n \to \infty$.

Theorem 12.6 *Let $A : X \to X$ be a compact linear operator on a Banach space X such that $I - A$ is injective. Assume that the sequence $A_n : X \to X$ of bounded linear operators is norm convergent, i.e., $\|A_n - A\| \to 0$, $n \to \infty$. Then for sufficiently large n the inverse operators $(I - A_n)^{-1} : X \to X$ exist and are uniformly bounded. For the solutions of the equations*

$$\varphi - A\varphi = f \quad \textit{and} \quad \varphi_n - A_n\varphi_n = f_n$$

we have an error estimate

$$\|\varphi_n - \varphi\| \le C\{\|(A_n - A)\varphi\| + \|f_n - f\|\} \tag{12.5}$$

for some constant C.

Proof. By the Riesz Theorem 12.2, the inverse $(I - A)^{-1} : X \to X$ exists and is bounded. Since $\|A_n - A\| \to 0$, $n \to \infty$, by Remark 3.25 we have $\|(I - A)^{-1}(A_n - A)\| \le q < 1$ for sufficiently large n. For these n, by the Neumann series Theorem 3.48, the inverse operators of

$$I - (I - A)^{-1}(A_n - A) = (I - A)^{-1}(I - A_n)$$

exist and are uniformly bounded by

$$\|[I - (I - A)^{-1}(A_n - A)]^{-1}\| \le \frac{1}{1 - q}\,.$$

But then $[I - (I - A)^{-1}(A - A_n)]^{-1}(I - A)^{-1}$ are the inverse operators of $I - A_n$ and they are uniformly bounded.

The error estimate follows from

$$(I - A_n)(\varphi_n - \varphi) = (A - A_n)\varphi + f_n - f$$

by the uniform boundedness of the inverse operators $(I - A_n)^{-1}$. □

In order to develop a similar analysis for the case where the sequence (A_n) is merely pointwise convergent, i.e., $A_n\varphi \to \varphi$, $n \to \infty$, for all φ, we will have to bridge the gap between norm and pointwise convergence. This goal will be achieved through the concept of collectively compact operator sequences and the following *uniform boundedness principle.*

Theorem 12.7 *Let the sequence $A_n : X \to Y$ of bounded linear operators mapping a Banach space X into a normed space Y be pointwise bounded;*

i.e., for each $\varphi \in X$ there exists a positive number C_φ depending on φ such that $\|A_n\varphi\| \le C_\varphi$ for all $n \in \mathbb{N}$. Then the sequence (A_n) is uniformly bounded; i.e., there exists some constant C such that $\|A_n\| \le C$ for all $n \in \mathbb{N}$.

Proof. In the first step, by an indirect proof we establish that positive constants M and ρ and an element $\psi \in X$ can be chosen such that

$$\|A_n\varphi\| \le M \tag{12.6}$$

for all $\varphi \in X$ with $\|\varphi - \psi\| \le \rho$ and all $n \in \mathbb{N}$. Assume that this is not possible. Then, by induction, we construct sequences (n_k) in $\mathbb{N}$, (ρ_k) in $\mathbb{R}$, and (φ_k) in X such that

$$\|A_{n_k}\varphi\| \ge k$$

for $k = 0, 1, 2, \dots$ and φ with $\|\varphi - \varphi_k\| \le \rho_k$ and

$$0 < \rho_k \le \frac{1}{2}\,\rho_{k-1}, \quad \|\varphi_k - \varphi_{k-1}\| \le \frac{1}{2}\,\rho_{k-1}$$

for $k = 1, 2, \dots$.

We initiate the induction by setting $n_0 = 1$, $\rho_0 = 1$, and $\varphi_0 = 0$. Assume that $n_k \in \mathbb{N}$, $\rho_k > 0$, and $\varphi_k \in X$ are given. Then there exist $n_{k+1} \in \mathbb{N}$ and $\varphi_{k+1} \in X$ satisfying $\|\varphi_{k+1} - \varphi_k\| \le \rho_k/2$ and $\|A_{n_{k+1}}\varphi_{k+1}\| \ge k+2$. Otherwise, we would have $\|A_n\varphi\| \le k+2$ for all $\varphi \in X$ with $\|\varphi - \varphi_k\| \le \rho_k/2$ and all $n \in \mathbb{N}$, and this contradicts our assumption. Set

$$\rho_{k+1} := \min\left(\frac{\rho_k}{2}, \frac{1}{\|A_{n_{k+1}}\|}\right) \le \frac{\rho_k}{2}\,.$$

Then for all $\varphi \in X$ with $\|\varphi - \varphi_{k+1}\| \le \rho_{k+1}$, by the triangle inequality we have

$$\|A_{n_{k+1}}\varphi\| \ge \|A_{n_{k+1}}\varphi_{k+1}\| - \|A_{n_{k+1}}(\varphi - \varphi_{k+1})\| \ge k+1,$$

since $\|A_{n_{k+1}}(\varphi - \varphi_{k+1})\| \le \|A_{n_{k+1}}\|\rho_{k+1} \le 1$.

For $j > k$, using the geometric series we have

$$\begin{aligned}\|\varphi_k - \varphi_j\| &\le \|\varphi_k - \varphi_{k+1}\| + \cdots + \|\varphi_{j-1} - \varphi_j\| \\ &\le \frac{1}{2}\,\rho_k + \cdots + \frac{1}{2}\,\rho_{j-1} \le \rho_k.\end{aligned}$$

Therefore, (φ_k) is a Cauchy sequence and converges to some element φ in the Banach space X. From $\|\varphi_k - \varphi_j\| \le \rho_k$ for all $j \ge k$ by passing to the limit $j \to \infty$ we see that $\|\varphi_k - \varphi\| \le \rho_k$ for all $k \in \mathbb{N}$. Therefore, we have $\|A_{n_k}\varphi\| \ge k$ for all $k \in \mathbb{N}$, which is a contradiction to the boundedness of the sequence $(A_n\varphi)$.

Now, in the second step, from the validity of (12.6) we deduce for each $\varphi \in X$ with $\|\varphi\| \leq 1$ and for all $n \in \mathbb{N}$ the estimate

$$\|A_n\varphi\| = \frac{1}{\rho}\,\|A_n(\rho\varphi + \psi) - A_n\psi\| \leq \frac{2M}{\rho}\,.$$

This completes the proof. □

Following Anselone [2], we introduce the concept of collectively compact operator sequences.

Definition 12.8 *A sequence $A_n : X \to Y$ of linear operators from a normed space X into a normed space Y is called* collectively compact *if each sequence from the set $\{A_n\varphi : \varphi \in X, \|\varphi\| \leq 1, n \in \mathbb{N}\}$ contains a convergent subsequence.*

Each operator A_n from a collectively compact sequence is compact.

Lemma 12.9 *Let X be a Banach space, let $A_n : X \to X$ be a collectively compact sequence, and let $B_n : X \to X$ be a pointwise convergent sequence with limit operator $B : X \to X$. Then*

$$\|(B_n - B)A_n\| \to 0, \quad n \to \infty. \tag{12.7}$$

Proof. Assume that (12.7) is not valid. Then there exist $\varepsilon_0 > 0$, a sequence (n_k) in $\mathbb{N}$ with $n_k \to \infty$, $k \to \infty$, and a sequence (φ_k) in X with $\|\varphi_k\| \leq 1$ such that

$$\|(B_{n_k} - B)A_{n_k}\varphi_k\| \geq \varepsilon_0, \quad k = 1, 2, \ldots. \tag{12.8}$$

Since the sequence (A_n) is collectively compact, there exists a subsequence such that

$$A_{n_{k(j)}}\varphi_{k(j)} \to \psi \in X, \quad j \to \infty. \tag{12.9}$$

Then we can estimate with the aid of the triangle inequality and Remark 3.25 to obtain

$$\begin{aligned} &\|(B_{n_{k(j)}} - B)A_{n_{k(j)}}\varphi_{k(j)}\| \\ &\leq \|(B_{n_{k(j)}} - B)\psi\| + \|B_{n_{k(j)}} - B\|\,\|A_{n_{k(j)}}\varphi_{k(j)} - \psi\|. \end{aligned} \tag{12.10}$$

The first term on the right-hand side of (12.10) tends to zero as $j \to \infty$, since the operator sequence (B_n) is pointwise convergent. The second term tends to zero as $j \to \infty$, since the operator sequence (B_n) is uniformly bounded by Theorem 12.7 and since we have the convergence (12.9). Therefore, passing to the limit $j \to \infty$ in (12.10) yields a contradiction to (12.8), and the proof is complete. □

Theorem 12.10 *Let $A : X \to X$ be a compact linear operator on a Banach space X such that $I - A$ is injective, and assume that the sequence $A_n : X \to X$ of linear operators is collectively compact and pointwise convergent; i.e., $A_n\varphi \to A\varphi$, $n \to \infty$, for all $\varphi \in X$. Then for sufficiently large n the inverse operators $(I - A_n)^{-1} : X \to X$ exist and are uniformly bounded. For the solutions of the equations*

$$\varphi - A\varphi = f \quad \textit{and} \quad \varphi_n - A_n\varphi_n = f_n$$

we have an error estimate

$$\|\varphi_n - \varphi\| \le C\{\|(A_n - A)\varphi\| + \|f_n - f\|\} \tag{12.11}$$

for some constant C.

Proof. By the Riesz Theorem 12.2, the inverse $(I - A)^{-1} : X \to X$ exists and is bounded. The identity

$$(I - A)^{-1} = I + (I - A)^{-1}A$$

suggests

$$M_n := I + (I - A)^{-1}A_n$$

as an approximate inverse for $I - A_n$. Elementary calculations yield

$$M_n(I - A_n) = I - S_n, \tag{12.12}$$

where

$$S_n := (I - A)^{-1}(A_n - A)A_n.$$

From Lemma 12.9 we conclude that $\|S_n\| \to 0$, $n \to \infty$. Hence for sufficiently large n we have $\|S_n\| \le q < 1$. For these n, by the Neumann series Theorem 3.48, the inverse operators $(I - S_n)^{-1}$ exist and are uniformly bounded by

$$\|(I - S_n)^{-1}\| \le \frac{1}{1 - q}\,.$$

Now (12.12) implies first that $I - A_n$ is injective, and therefore, since A_n is compact, by Theorem 12.1 the inverse $(I - A_n)^{-1}$ exists. Then (12.12) also yields $(I - A_n)^{-1} = (I - S_n)^{-1}M_n$, whence uniform boundedness follows, since the operators M_n are uniformly bounded by Theorem 12.7. The error estimate (12.11) is proven as in Theorem 12.6. □

Note that both error estimates (12.5) and (12.11) show that the accuracy of the approximate solution essentially depends on how well $A_n\varphi$ approximates $A\varphi$ for the exact solution φ.

12.3 Nyström's Method

Recalling Chapter 9, we choose a convergent sequence

$$Q_n(g) = \sum_{k=0}^{n} a_k^{(n)} g(x_k^{(n)})$$

of quadrature formulae for the integral

$$Q(g) = \int_a^b g(x)\,dx$$

with quadrature points $x_0^{(n)}, \dots, x_n^{(n)} \in [a,b]$ and real quadrature weights $a_0^{(n)}, \dots, a_n^{(n)}$. For convenience we write $x_0, \dots, x_n$ instead of $x_0^{(n)}, \dots, x_n^{(n)}$, and $a_0, \dots, a_n$ instead of $a_0^{(n)}, \dots, a_n^{(n)}$. We approximate the integral operator

$$(A\varphi)(x) = \int_a^b K(x,y)\varphi(y)\,dy, \quad x \in [a,b],$$

with continuous kernel K by a sequence of numerical integration operators

$$(A_n\varphi)(x) := \sum_{k=0}^{n} a_k K(x, x_k)\varphi(x_k), \quad x \in [a,b];$$

i.e., we apply the quadrature formulae for $g = K(x,\cdot)\varphi$. Then the solution to the integral equation of the second kind

$$\varphi - A\varphi = f$$

is approximated by the solution of

$$\varphi_n - A_n\varphi_n = f,$$

which reduces to solving a finite-dimensional linear system.

Theorem 12.11 *Let φ_n be a solution of*

$$\varphi_n(x) - \sum_{k=0}^{n} a_k K(x, x_k)\varphi_n(x_k) = f(x), \quad x \in [a,b]. \tag{12.13}$$

Then the values $\varphi_j^{(n)} := \varphi_n(x_j)$, $j = 0, \dots, n$, at the quadrature points satisfy the linear system

$$\varphi_j^{(n)} - \sum_{k=0}^{n} a_k K(x_j, x_k)\varphi_k^{(n)} = f(x_j), \quad j = 0, \dots, n. \tag{12.14}$$

Conversely, let $\varphi_j^{(n)}$, $j = 0, \ldots, n$, be a solution of the system (12.14). Then the function φ_n defined by

$$\varphi_n(x) := f(x) + \sum_{k=0}^{n} a_k K(x, x_k)\varphi_k^{(n)}, \quad x \in [a, b], \tag{12.15}$$

solves equation (12.13).

Proof. The first statement is trivial. For a solution $\varphi_j^{(n)}$, $j = 0, \ldots, n$, of the system (12.14) the function φ_n defined by (12.15) has values

$$\varphi_n(x_j) = f(x_j) + \sum_{k=0}^{n} a_k K(x_j, x_k)\varphi_k^{(n)} = \varphi_j^{(n)}, \quad j = 0, \ldots, n.$$

Inserting this into (12.15), we see that φ_n satisfies (12.13). □

The formula (12.15) may be viewed as a natural interpolation of the values $\varphi_j^{(n)}$, $j = 0, \ldots, n$, at the quadrature points to obtain the approximating function φ_n. It was introduced by Nyström in 1930.

For convenience we note the following analogue of Theorem 12.5.

Theorem 12.12 *The norm of the quadrature operators A_n is given by*

$$\|A_n\|_\infty = \max_{a \le x \le b} \sum_{k=0}^{n} |a_k K(x, x_k)|. \tag{12.16}$$

Proof. For each $\varphi \in C[a, b]$ with $\|\varphi\|_\infty \le 1$ we have

$$\|A_n \varphi\|_\infty \le \max_{a \le x \le b} \sum_{k=0}^{n} |a_k K(x, x_k)|,$$

and therefore $\|A_n\|_\infty$ is smaller than or equal to the right-hand side of (12.16). Let $z \in [a, b]$ be such that

$$\sum_{k=0}^{n} |a_k K(z, x_k)| = \max_{a \le x \le b} \sum_{k=0}^{n} |a_k K(x, x_k)|$$

and choose $\psi \in C[a, b]$ with $\|\psi\|_\infty = 1$ and

$$a_k K(z, x_k)\psi(x_k) = |a_k K(z, x_k)|, \quad k = 0, \ldots, n.$$

Then

$$\|A_n\|_\infty \ge \|A_n \psi\|_\infty \ge |(A_n \psi)(z)| = \sum_{k=0}^{n} |a_k K(z, x_k)|,$$

and (12.16) is proven. □

The error analysis will be based on the following theorem.

Theorem 12.13 *Assume the quadrature formulae (Q_n) to be convergent. Then the sequence (A_n) is collectively compact and pointwise convergent (i.e., $A_n\varphi \to A\varphi$, $n \to \infty$, for all $\varphi \in C[a,b]$) but not norm convergent.*

Proof. Since the quadrature formulae (Q_n) are assumed to be convergent, by (9.13) and the uniform boundedness principle Theorem 12.7 there exists a constant C such that the weights satisfy

$$\sum_{k=0}^{n} |a_k^{(n)}| \leq C$$

for all $n \in \mathbb{N}$ (see Theorem 9.10). Then we can estimate

$$\|A_n\varphi\|_\infty \leq C \max_{a\leq x,y\leq b} |K(x,y)|\, \|\varphi\|_\infty \tag{12.17}$$

and

$$|(A_n\varphi)(x_1) - (A_n\varphi)(x_2)| \leq C \max_{a\leq y\leq b} |K(x_1,y) - K(x_2,y)|\, \|\varphi\|_\infty \tag{12.18}$$

for all $x_1, x_2 \in [a,b]$. From (12.17) and (12.18) we see that

$$\{A_n\varphi : \varphi \in C[a,b],\ \|\varphi\|_\infty \leq 1,\ n \in \mathbb{N}\}$$

is bounded and equicontinuous because the kernel K is uniformly continuous on $[a,b] \times [a,b]$. Therefore, by the Arzelà–Ascoli Theorem 12.3 the sequence (A_n) is collectively compact.

Since the quadrature is convergent, for fixed $\varphi \in C[a,b]$ the sequence $(A_n\varphi)$ is pointwise convergent; i.e., $(A_n\varphi)(x) \to (A\varphi)(x)$, $n \to \infty$, for all $x \in [a,b]$. As a consequence of (12.18), the sequence $(A_n\varphi)$ is equicontinuous. Hence it is uniformly convergent: $\|A_n\varphi - A\varphi\|_\infty \to 0$, $n \to \infty$. That is, we have pointwise convergence: $A_n\varphi \to A\varphi$, $n \to \infty$, for all $\varphi \in C[a,b]$ (see Problem 12.7).

For $\varepsilon > 0$ choose a function $\psi_\varepsilon \in C[a,b]$ with $0 \leq \psi_\varepsilon(x) \leq 1$ for all $x \in [a,b]$ such that $\psi_\varepsilon(x) = 1$ if $\min_{j=0,\dots,n} |x - x_j| \geq \varepsilon$ and $\psi_\varepsilon(x_j) = 0$, $j = 0,\dots,n$. Then

$$\|A(\varphi\psi_\varepsilon) - A\varphi\|_\infty \leq \max_{x,y\in[a,b]} |K(x,y)| \int_a^b \{1 - \psi_\varepsilon(y)\}\, dy \to 0, \quad \varepsilon \to 0,$$

for all $\varphi \in C[a,b]$ with $\|\varphi\|_\infty = 1$. Using this result, we derive

$$\begin{aligned}\|A - A_n\|_\infty &= \sup_{\|\varphi\|_\infty=1} \|(A - A_n)\varphi\|_\infty \geq \sup_{\|\varphi\|_\infty=1} \sup_{\varepsilon>0} \|(A - A_n)(\varphi\psi_\varepsilon)\|_\infty \\ &= \sup_{\|\varphi\|_\infty=1} \sup_{\varepsilon>0} \|A(\varphi\psi_\varepsilon)\|_\infty \geq \sup_{\|\varphi\|_\infty=1} \|A\varphi\|_\infty = \|A\|_\infty,\end{aligned}$$

whence we see that the sequence (A_n) cannot be norm convergent. □

Theorem 12.13 enables us to apply the approximation theory of Theorem 12.10. For the discussion of the error based on the estimate (12.11) we need the norm $\|A\varphi - A_n\varphi\|_\infty$. It can be expressed in terms of the error for the corresponding numerical quadrature by

$$\|A\varphi - A_n\varphi\|_\infty = \max_{a\le x\le b}\left|\int_a^b K(x,y)\varphi(y)\,dy - \sum_{k=0}^n a_k K(x,x_k)\varphi(x_k)\right|$$

and requires a uniform estimate for the error of the quadrature applied to the integration of $K(x,\cdot)\varphi$. Therefore, from the error estimate (12.11), it follows that under suitable regularity assumptions on the kernel K and the exact solution φ, the convergence order of the underlying quadrature formulae carries over to the convergence order of the approximate solutions to the integral equation. We illustrate this by the case of the trapezoidal rule. Under the assumption $\varphi \in C^2[a,b]$ and $K \in C^2([a,b]\times[a,b])$, by Theorem 9.7, we can estimate

$$\|A\varphi - A_n\varphi\|_\infty \le \frac{1}{12}\,h^2(b-a)\max_{a\le x,y\le b}\left|\frac{\partial^2}{\partial y^2}\,[K(x,y)\varphi(y)]\right|.$$

Example 12.14 Consider the integral equation

$$\varphi(x) - \frac{1}{2}\int_0^1 (x+1)e^{-xy}\varphi(y)dy = e^{-x} - \frac{1}{2} + \frac{1}{2}\,e^{-(x+1)},\quad 0\le x\le 1, \tag{12.19}$$

with exact solution $\varphi(x) = e^{-x}$. For its kernel we have

$$\max_{0\le x\le 1}\int_0^1 \frac{1}{2}\,(x+1)e^{-xy}dy = \sup_{0<x\le 1}\frac{x+1}{2x}\,(1-e^{-x}) < 1.$$

Therefore, by the Neumann series Theorem 3.48 and the operator norm (12.4), equation (12.19) is uniquely solvable.

We use the (composite) trapezoidal rule for approximately solving the integral equation (12.19) by the Nyström method. Table 12.1 gives the difference between the exact and approximate solutions and clearly shows the expected convergence rate $O(h^2)$.

TABLE 12.1. Numerical solution of (12.19) by the trapezoidal rule

n	$x=0$	$x=0.25$	$x=0.5$	$x=0.75$	$x=1$
4	0.007146	0.008878	0.010816	0.013007	0.015479
8	0.001788	0.002224	0.002711	0.003261	0.003882
16	0.000447	0.000556	0.000678	0.000816	0.000971
32	0.000112	0.000139	0.000170	0.000204	0.000243

We now use the (composite) Simpson's rule for the integral equation (12.19). The numerical results in Table 12.2 show the convergence order $O(h^4)$, which we expect from the error estimate (12.11) and the convergence order for Simpson's rule from Theorem 9.8. □

TABLE 12.2. Numerical solution of (12.19) by Simpson's rule

n	$x = 0$	$x = 0.25$	$x = 0.5$	$x = 0.75$	$x = 1$
4	0.00006652	0.00008311	0.00010905	0.00015046	0.00021416
8	0.00000422	0.00000527	0.00000692	0.00000956	0.00001366
16	0.00000026	0.00000033	0.00000043	0.00000060	0.00000086

After comparing Tables 12.1 and 12.2, we wish to emphasize the major advantage of Nyström's method over other methods like the collocation method, which we will discuss in the next section. The matrix and the right-hand side of the linear system (12.14) are obtained by just evaluating the kernel K and the given function f at the quadrature points. Therefore, without any further computational effort we can improve considerably on the approximations by choosing a more accurate numerical quadrature formula.

In the next example we consider an integral equation with a periodic kernel and a periodic solution.

Example 12.15 Consider the integral equation

$$\varphi(t) + \frac{ab}{\pi} \int_0^{2\pi} \frac{\varphi(\tau)d\tau}{a^2 + b^2 - (a^2 - b^2)\cos(t + \tau)} = f(t), \quad 0 \le t \le 2\pi, \tag{12.20}$$

where $a \ge b > 0$. This integral equation arises from the solution of the Dirichlet problem for the Laplace equation in an ellipse with semiaxis a and b (see [39]). Any solution φ to the homogeneous form of equation (12.20) clearly must be a 2π-periodic analytic function, since the kernel is a 2π-periodic analytic function with respect to the variable t. Hence, we can expand φ into a uniformly convergent Fourier series

$$\varphi(t) = \sum_{n=0}^{\infty} \alpha_n \cos nt + \sum_{n=1}^{\infty} \beta_n \sin nt.$$

Inserting this into the homogeneous integral equation and using the integrals (see Problem 12.10)

$$\frac{ab}{\pi} \int_0^{2\pi} \frac{e^{in\tau}d\tau}{(a^2 + b^2) - (a^2 - b^2)\cos(t + \tau)} = \left(\frac{a-b}{a+b}\right)^n e^{-int} \tag{12.21}$$

for $n = 0, 1, 2, \ldots$, it follows that

$$\alpha_n \left[1 + \left(\frac{a-b}{a+b}\right)^n\right] = \beta_n \left[1 - \left(\frac{a-b}{a+b}\right)^n\right] = 0$$

for $n = 0, 1, 2, \ldots$. Hence, $\alpha_n = \beta_n = 0$ for $n = 0, 1, 2, \ldots$, and therefore $\varphi = 0$. Now the Riesz Theorem 12.2 implies that the integral equation (12.20) is uniquely solvable for each right-hand side f.

We numerically want to solve (12.20) in the case where the unique solution is given by

$$\varphi(t) = e^{\cos t} \cos(\sin t), \quad 0 \le t \le 2\pi.$$

Using the integrals (12.21), it can be seen that the right-hand side becomes

$$f(t) = \varphi(t) + e^{c \cos t} \cos(c \sin t), \quad 0 \le t \le 2\pi,$$

where $c = (a - b)/(a + b)$.

Since we are dealing with periodic analytic functions, we use the rectangular rule. From Theorem 9.28 we expect an exponentially decreasing error behavior, which is exhibited by the numerical results in Table 12.3 giving the difference between the exact and approximate solutions. Doubling the number of quadrature points doubles the number of correct digits in the approximate solution.

TABLE 12.3. Nyström method for equation (12.20)

	n	$t = 0$	$t = \pi/2$	$t = \pi$
	4	−0.15350443	0.01354412	−0.00636277
$a = 1$	8	−0.00281745	0.00009601	−0.00004247
$b = 0.5$	16	−0.00000044	0.00000001	−0.00000001
	4	−0.69224130	−0.06117951	−0.06216587
$a = 1$	8	−0.15017166	−0.00971695	−0.01174302
$b = 0.2$	16	−0.00602633	−0.00036043	−0.00045498
	32	−0.00000919	−0.00000055	−0.00000069

The actual size of the error, i.e., the constant factor in the exponential decay, depends on the parameters a and b, which describe the location of the singularities of the integrands in the complex plane; i.e., they determine the width of the strip of the complex plane into which the kernel can be extended as a holomorphic function.

Note that for periodic analytic functions the rectangular rule generally yields better approximations than Simpson's rule (see Problem 9.12). □

We confine ourselves to these few examples for the application of the Nyström method. For a greater variety the reader is referred to [1, 3, 6, 19, 25, 30, 39, 49].

With the aid of appropriately chosen quadrature formulae, which take care of the singularity by a weighted product rule, the Nyström method can also be successfully applied to weakly singular integral equations of the second kind (see [39]).

12.4 The Collocation Method

The *collocation method* for approximately solving an equation of the second kind

$$\varphi - A\varphi = f \tag{12.22}$$

consists in seeking an approximate solution from a finite-dimensional subspace by requiring that the equation (12.22) be satisfied at only a finite number of so-called *collocation points*. Assume that $A : C[a,b] \to C[a,b]$ is a bounded linear operator and let $X_n = \text{span}\{u_0^{(n)}, \dots, u_n^{(n)}\} \subset C[a,b]$ denote a sequence of subspaces with $\dim X_n = n+1$. Choose $n+1$ points $a \le x_0^{(n)} < \cdots < x_n^{(n)} \le b$ such that the interpolation at these grid points with respect to the subspace X_n is uniquely solvable. Typical examples for the choice of X_n are polynomials, trigonometric polynomials, and splines (see also Problem 8.1). For convenience we will again write $x_0, \dots, x_n$ instead of $x_0^{(n)}, \dots, x_n^{(n)}$, and $u_0, \dots, u_n$ instead of $u_0^{(n)}, \dots, u_n^{(n)}$. By $L_n : C[a,b] \to X_n$ we denote the operator that maps the function $f \in C[a,b]$ into its uniquely determined interpolating function $L_n f \in X_n$ with the property

$$(L_n f)(x_j) = f(x_j), \quad j = 0, \dots, n.$$

Representing L_n in terms of the Lagrange basis, i.e., in terms of the uniquely determined functions $\ell_0, \dots, \ell_n \in X_n$ with the interpolation property

$$\ell_k(x_j) = \delta_{jk}, \quad j, k = 0, \dots, n,$$

in the form

$$L_n f = \sum_{k=0}^{n} f(x_k)\ell_k \tag{12.23}$$

it can be seen that the operator $L_n : C[a,b] \to X_n$ is linear and bounded (with respect to the maximum norm). Moreover, since $L_n f = f$ for all $f \in X_n$, the interpolation operator is a projection operator; i.e., $L_n^2 = L_n$ (see p. 157 and Problem 8.4)

The collocation method approximates the solution of (12.22) by an element $\varphi_n \in X_n$ satisfying

$$\varphi_n(x_j) - (A\varphi_n)(x_j) = f(x_j), \quad j = 0, \dots, n. \tag{12.24}$$

We express φ_n as a linear combination

$$\varphi_n = \sum_{k=0}^{n} \gamma_k u_k$$

and immediately see that equation (12.24) is equivalent to the linear system

$$\sum_{k=0}^{n} \gamma_k \{u_k(x_j) - (Au_k)(x_j)\} = f(x_j), \quad j = 0, \dots, n, \tag{12.25}$$

for the coefficients $\gamma_0, \dots, \gamma_n$. If we use the Lagrange basis for X_n and write

$$\varphi_n = \sum_{k=0}^{n} \gamma_k \ell_k,$$

then of course $\gamma_j = \varphi_n(x_j)$, $j = 0, \dots, n$, and the system (12.25) becomes

$$\gamma_j - \sum_{k=0}^{n} \gamma_k (A\ell_k)(x_j) = f(x_j), \quad j = 0, \dots, n. \tag{12.26}$$

From the systems (12.25) and (12.26) it is obvious that the collocation method is only semidiscrete, since in general, additional approximations are needed in order to compute the matrix entries $(Au_k)(x_j)$ or $(A\ell_k)(x_j)$.

The collocation method can be interpreted as a *projection method*; i.e., since the interpolating function is uniquely determined by its values at the interpolation points, equation (12.24) is equivalent to

$$\varphi_n - L_n A\varphi_n = L_n f. \tag{12.27}$$

This equation can be considered as an equation in the whole space $C[a, b]$ because any solution $\varphi_n = L_n A\varphi_n + L_n f$ automatically belongs to X_n. Hence, our general error and convergence results for operator equations of the second kind can be applied to the collocation method.

Theorem 12.16 *Let $A : C[a, b] \to C[a, b]$ be a compact linear operator such that $I - A$ is injective, and assume that the interpolation operators $L_n : C[a, b] \to X_n$ satisfy $\|L_n A - A\|_\infty \to 0$, $n \to \infty$. Then, for sufficiently large n, the approximate equation (12.27) is uniquely solvable for all $f \in C[a, b]$, and we have the error estimate*

$$\|\varphi_n - \varphi\|_\infty \le C\|L_n\varphi - \varphi\|_\infty \tag{12.28}$$

for some positive constant C depending on A.

Proof. From Theorem 12.6 applied to $A_n = L_n A$, we conclude that for all sufficiently large n the inverse operators $(I - L_n A)^{-1}$ exist and are uniformly bounded. To verify the error bound, we apply the interpolation operator L_n to (12.22) and get

$$\varphi - L_n A\varphi = L_n f + \varphi - L_n \varphi.$$

Subtracting this from (12.27) we find

$$(I - L_n A)(\varphi_n - \varphi) = L_n\varphi - \varphi,$$

whence the estimate (12.28) follows. □

Corollary 12.17 *Let $A : C[a,b] \to C[a,b]$ be a compact linear operator such that $I - A$ is injective, and assume that the interpolation operators $L_n : C[a,b] \to X_n$ are pointwise convergent; i.e., $L_n\varphi \to \varphi$, $n \to \infty$, for all $\varphi \in C[a,b]$. Then, for sufficiently large n, the approximate equation (12.27) is uniquely solvable for all $f \in C[a,b]$, and the estimate (12.28) holds.*

Proof. By Lemma 12.9 the pointwise convergence of the interpolation operators L_n and the compactness of A imply that $\|L_n A - A\|_\infty \to 0$, $n \to \infty$. Now the statement follows from the preceding theorem. □

We note that the collocation method may of course also be applied in function spaces other than the space $C[a,b]$.

We proceed by considering the collocation method for integral equations of the second kind

$$\varphi(x) - \int_a^b K(x,y)\varphi(y)\,dy = f(x), \quad x \in [a,b], \tag{12.29}$$

with continuous kernel K. Using the interpolation operator, in this case we can rewrite the collocation equation (12.26) in the form

$$\varphi_n(x) - \int_a^b [L_n K(\cdot\,,y)](x)\varphi_n(y)\,dy = (L_n f)(x), \quad x \in [a,b], \tag{12.30}$$

and the systems (12.25) and (12.26) become

$$\sum_{k=0}^n \gamma_k \left\{ u_k(x_j) - \int_a^b K(x_j,y)u_k(y)\,dy \right\} = f(x_j), \quad j = 0,\dots,n, \tag{12.31}$$

and

$$\gamma_j - \sum_{k=0}^n \gamma_k \int_a^b K(x_j,y)\ell_k(y)\,dy = f(x_j), \quad j = 0,\dots,n, \tag{12.32}$$

respectively. There exists a broad variety of collocation methods corresponding to various choices for the subspaces X_n, for the basis functions

$u_0, \dots, u_n$, and for the collocation points $x_0, \dots, x_n$. We briefly discuss two possibilities, based on linear splines and on trigonometric polynomials.

First we consider piecewise linear interpolation. Let $x_j = a + jh$, $j = 0, \dots, n$, denote an equidistant subdivision with step size $h = (b-a)/n$ and let X_n be the space of continuous functions on $[a, b]$ whose restrictions on each of the subintervals $[x_{j-1}, x_j]$, $j = 1, \dots, n$, coincide with a linear function. As in Section 11.5, the Lagrange basis is given by

$$\ell_k(x) := \begin{cases} \dfrac{1}{h}\,(x - x_{k-1}), & x \in [x_{k-1}, x_k], \\[2ex] \dfrac{1}{h}\,(x_{k+1} - x), & x \in [x_k, x_{k+1}], \\[2ex] 0, & x \notin [x_{k-1}, x_{k+1}], \end{cases}$$

for $k = 0, \dots, n$. Since for piecewise linear interpolation we have that

$$\|L_n f\|_\infty \le \max_{j=0,\dots,n} |f(x_j)| \le \|f\|_\infty,$$

with equality holding if f is constant, we observe that $\|L_n\|_\infty = 1$ for the corresponding interpolation operator L_n. Here, we have pointwise convergence $L_n\varphi \to \varphi$, $n \to \infty$. This can be seen from the error estimate (8.9) and the Weierstrass approximation theorem, analogous to the proof of the Szegő Theorem 9.10. Therefore, in this case Corollary 12.17 applies, and we can state the following result.

Theorem 12.18 *The collocation method with linear splines converges for integral equations of the second kind with continuous kernels.*

Provided that the exact solution of the integral equation is twice continuously differentiable, then from the error estimate (8.9) for linear interpolation and Corollary 12.17 we derive an error estimate of the form

$$\|\varphi_n - \varphi\|_\infty \le C\|\varphi''\|_\infty h^2$$

for the linear spline collocation approximate solution φ_n. Here, C denotes some constant depending on the kernel K.

In general, in most practical problems the evaluation of the matrix entries in (12.32) will require a numerical quadrature for integrals of the form $\int_a^b K(x_j, y)\ell_k(y)\,dy$. To be consistent with our approximations, we replace $K(x_j, \cdot)$ by its piecewise linear interpolation; i.e., we approximate

$$\int_a^b K(x_j, y)\ell_k(y)\,dy \approx \sum_{i=0}^n K(x_j, x_i) \int_a^b \ell_i(y)\ell_k(y)\,dy$$

for $j, k = 0, \dots, n$. Straightforward calculations yield the tridiagonal matrix

$$W = \frac{h}{6} \begin{pmatrix} 2 & 1 & & & & \\ 1 & 4 & 1 & & & \\ & 1 & 4 & 1 & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & 1 & 4 & 1 \\ & & & & 1 & 2 \end{pmatrix}$$

for the weights $w_{ik} = \int_a^b \ell_i(y)\ell_k(y)\,dy$.

We now investigate the influence of these approximations on the error analysis. We interpret the solution of the system (12.32) with the approximate values for the coefficients as the solution $\tilde{\varphi}_n$ of an additional approximate equation

$$\tilde{\varphi}_n - A_n\tilde{\varphi}_n = L_n f, \tag{12.33}$$

namely of the collocation equation

$$\tilde{\varphi}_n(x) - \int_a^b [L_n K_n(\cdot\,, y)](x)\tilde{\varphi}_n(y)\,dy = (L_n f)(x), \quad a \le x \le b,$$

with

$$K_n(x, y) := \sum_{i=0}^{n} K(x, x_i)\ell_i(y);$$

i.e., $K_n(x, y) = [L_n K_n(x, \cdot)](y)$ interpolates K with respect to the second variable. We assume that the kernel K is twice continuously differentiable on $[a, b] \times [a, b]$. Then, using the error estimate (8.9), we have

$$|K(x, y) - K_n(x, y)| \le \frac{h^2}{8} \left\| \frac{\partial^2 K}{\partial y^2} \right\|_\infty$$

for all $a \le x, y \le b$. Writing

$$K_n(x, y) - [L_n K_n(\cdot\,, y)](x) = L_n\{K(x, \cdot) - [L_n K_n(\cdot\,, \cdot)](x)\}(y)$$

and using the fact that for the piecewise linear spline interpolation we have $\|L_n\|_\infty = 1$, from (8.9) we obtain

$$|K_n(x, y) - [L_n K_n(\cdot\,, y)](x)| \le \frac{h^2}{8} \left\| \frac{\partial^2 K}{\partial x^2} \right\|_\infty$$

for all $a \le x, y \le b$. Hence, in view of (12.4), for the integral operator A_n with kernel K_n we have $\|A_n - A\|_\infty = O(h^2)$. When f is twice continuously differentiable, we also have $\|L_n f - f\|_\infty = O(h^2)$. Therefore, from Theorem 12.6 we can now conclude that the approximate equation (12.33) is uniquely solvable for sufficiently large n and that for the approximate solution we have an error estimate $\|\tilde{\varphi}_n - \varphi\|_\infty = O(h^2)$. Therefore, the fully discrete approximation still is of order $O(h^2)$.

Example 12.19 Consider the integral equation (12.19) of Example 12.14. Table 12.4 gives the error between the exact solution and the fully discrete collocation approximation with linear splines. It clearly exhibits the error behavior $O(h^2)$. □

TABLE 12.4. Numerical results for spline collocation

n	$x = 0$	$x = 0.25$	$x = 0.5$	$x = 0.75$	$x = 1$
4	0.004808	0.005430	0.006178	0.007128	0.008331
8	0.001199	0.001354	0.001541	0.001778	0.002078
16	0.000300	0.000338	0.000385	0.000444	0.000519
32	0.000075	0.000085	0.000096	0.000111	0.000130

We note that in principle, a collocation method with error $O(h^4)$ can be obtained from cubic spline interpolation (see Theorem 8.34). However, the numerical implementation is much more involved. This again illustrates that the Nyström method is more practical, since there it is quite easy to change the order from $O(h^2)$ to $O(h^4)$ by simply replacing the weights of the trapezoidal rule by those of Simpson's rule.

We proceed by discussing the collocation method based on trigonometric interpolation with equidistant knots $t_j = j\pi/n$, $j = 0, \dots, 2n-1$. First, we establish a convergence result for the trigonometric interpolation of differentiable functions (see Problem 8.12).

Lemma 12.20 *Let $f \in C^1[0, 2\pi]$. Then for the remainder in trigonometric interpolation we have*

$$||L_n f - f||_\infty \leq c_n ||f'||_2, \tag{12.34}$$

where $c_n \to 0$, $n \to \infty$.

Proof. Consider the trigonometric monomials $f_m(t) = e^{imt}$ and write $m = (2k+1)n + q$ with $k \in \mathbb{Z}$ and $0 \leq q < 2n$. Since $f_m(t_j) = f_{q-n}(t_j)$ for $j = 0, \dots, 2n-1$, the trigonometric interpolation polynomials for f_m and f_{q-n} coincide. Therefore, we have

$$||L_n f_m - f_m||_\infty \leq 2$$

for all $|m| \geq n$. Since f is continuously differentiable, we can expand it into a uniformly convergent Fourier series (see Problem 12.14)

$$f = \sum_{m=-\infty}^{\infty} a_m f_m.$$

From the relation

$$\int_0^{2\pi} f'(t)e^{-imt}dt = im\int_0^{2\pi} f(t)e^{-imt}dt = 2\pi i m a_m$$

for the Fourier coefficients it follows that

$$\|f'\|_2^2 = \int_0^{2\pi} |f'(t)|^2 dt = 2\pi \sum_{m=-\infty}^{\infty} m^2 |a_m|^2.$$

Using this identity and the Cauchy–Schwarz inequality, we derive

$$\|L_n f - f\|_\infty^2 \le 4\left\{\sum_{|m|=n}^{\infty} |a_m|\right\}^2 \le \frac{4}{\pi}\,\|f'\|_2^2 \sum_{m=n}^{\infty} \frac{1}{m^2}\,.$$

This implies (12.34). □

Now, consider an integral equation of the second kind with 2π-periodic continuously differentiable kernel K and right-hand side f. The corresponding integral operator A maps $C[0,2\pi]$ into $C^1[0,2\pi]$ and satisfies $\|(A\varphi)'\|_2 \le M\|\varphi\|_\infty$, where $M = \sqrt{2\pi}\,\|\partial K/\partial t\|_\infty$. Therefore, making use of (12.34), we find

$$\|L_n A\varphi - A\varphi\|_\infty \le c_n \|(A\varphi)'\|_2 \le c_n M\|\varphi\|_\infty$$

for all $\varphi \in C[0,2\pi]$. Hence, $\|L_n A - A\|_\infty \le c_n M \to 0$, $n \to \infty$, and Theorem 12.16 can be applied to obtain the following result.

Theorem 12.21 *The collocation method with trigonometric polynomials converges for integral equations of the second kind with continuously differentiable periodic kernels and right-hand sides.*

One possibility for the implementation of the collocation method is to use the trigonometric monomials as basis functions. Then the integrals $\int_0^{2\pi} K(t_j,\tau)e^{ik\tau}d\tau$ have to be integrated numerically. Replacing the kernel by its trigonometric interpolation leads to the quadrature formula

$$\int_0^{2\pi} K(t_j,\tau)e^{ik\tau}d\tau \approx \frac{\pi}{n}\sum_{m=0}^{2n-1} K(t_j,t_m)e^{ikt_m}$$

for $j = 0,\dots,2n-1$. Using fast Fourier transform techniques (see Section 8.2) these quadratures can be carried out very rapidly. A second, even more efficient, possibility is to use the Lagrange basis

$$\ell_k(t) = \frac{1}{2n}\left\{1 + 2\sum_{m=1}^{n-1}\cos m(t-t_k) + \cos n(t-t_k)\right\} \qquad (12.35)$$

for $k = 0, \dots, 2n-1$ which can be derived from Theorem 8.25 (see Problem 12.13).

For the evaluation of the matrix coefficients $\int_0^{2\pi} K(t_j,\tau)\ell_k(\tau)\,d\tau$ we proceed analogously to the preceding case of linear splines. We approximate these integrals by replacing $K(t_j,\cdot)$ by its trigonometric interpolation polynomial, i.e., we approximate

$$\int_0^{2\pi} K(t_j,\tau)\ell_k(\tau)\,d\tau \approx \sum_{m=0}^{2n-1} K(t_j,t_m)\int_0^{2\pi} \ell_m(\tau)\ell_k(\tau)\,d\tau$$

for $j,k = 0,\dots,2n-1$. Using (12.35), elementary integrations yield (see Problem 12.13)

$$\int_0^{2\pi} \ell_m(\tau)\ell_k(\tau)\,d\tau = \frac{\pi}{n}\,\delta_{mk} - (-1)^{m-k}\,\frac{\pi}{4n^2}\,, \tag{12.36}$$

for $m,k = 0,\dots,2n-1$. Note that despite the global nature of the trigonometric interpolation and its Lagrange basis, due to the simple structure of the weights (12.36) in the quadrature rule, the computation of the matrix elements is not too costly. The only additional computational effort besides the kernel evaluation is the computation of the row sums

$$\sum_{m=0}^{2n-1}(-1)^m K(t_j,t_m)$$

for $j = 0,\dots,2n-1$. We omit the analysis of the additional error in the fully discrete method caused by the numerical quadrature.

Example 12.22 For the integral equation (12.20) from Example 12.15, Table 12.5 gives the error between the exact solution and the collocation approximation.

TABLE 12.5. Collocation method for equation (12.20)

	n	$t=0$	$t=\pi/2$	$t=\pi$
	4	−0.10752855	−0.03243176	0.03961310
$a=1$	8	−0.00231537	0.00059809	0.00045961
$b=0.5$	16	−0.00000044	0.00000002	−0.00000000
	4	−0.56984945	−0.18357135	0.06022598
$a=1$	8	−0.14414257	−0.00368787	−0.00571394
$b=0.2$	16	−0.00602543	−0.00035953	−0.00045408
	32	−0.00000919	−0.00000055	−0.00000069

Again we have exponential convergence, as is to be expected from the estimate (12.28) and the error analysis for the trigonometric interpolation for analytic functions [38]. □

In general, the fully discrete implementation of the collocation method as described by our two examples can be used in all situations where the required numerical quadratures for the matrix elements can be carried out in closed form for the chosen approximating subspace and collocation points. In all these cases, of course, the quadrature formulae that are required for the related Nyström method will also be available. Because the approximation order for both methods usually will be the same, Nyström's method is preferable, since it requires the least computational effort for evaluating the matrix elements. However, the situation changes in cases where no straightforward quadrature rules for the application of Nyström's method are available.

Again, for a greater variety of collocation methods the reader is referred to [1, 3, 6, 19, 25, 30, 39, 49].

12.5 Stability

For finite-dimensional approximations of a given operator equation we have to distinguish three condition numbers, namely, the condition numbers of the original operator and of the approximating operator as mappings in the underlying normed spaces, and the condition number of the linear system for the actual numerical solution. This latter system we can influence, for example in the collocation method by the choice of the basis for the approximating subspaces.

Consider an equation of the second kind $\varphi - A\varphi = f$ in a Banach space X and approximating equations $\varphi_n - A_n\varphi_n = f_n$ under the assumptions of Theorem 12.6, i.e., norm convergence, or of Theorem 12.10, i.e., collective compactness and pointwise convergence. Then, recalling Definition 5.2 of the condition number, from Theorems 12.6 and 12.10 it follows that the condition numbers $\operatorname{cond}(I - A_n)$ are uniformly bounded. Hence, for the condition of the approximating scheme, we mainly have to be concerned with the condition of the linear system for the actual computation of the solution of $\varphi_n - A_n\varphi_n = f_n$.

For the discussion of the condition number for the Nyström method we recall the linear system (12.14) and denote by $\tilde{A}_n$ the matrix with the entries $a_k K(x_j, x_k)$. We introduce operators $R_n : C[a,b] \to \mathbb{R}^{n+1}$ by

$$R_n : f \mapsto (f(x_0), \dots, f(x_n))^T, \quad f \in C[a,b],$$

and $M_n : \mathbb{R}^{n+1} \to C[a,b]$, where $M_n\Phi$ is the piecewise linear interpolation with $(M_n\Phi)(x_j) = \Phi_j$, $j = 0, \dots, n$, for $\Phi = (\Phi_0, \dots, \Phi_n)^T$. (If $a < x_0$, we

set $(M_n\Phi)(x) = \Phi_0$ for $a \le x \le x_0$; and if $x_n < b$, we set $(M_n\Phi)(x) = \Phi_n$ for $x_n \le x \le b$.) Then clearly, $\|R_n\|_\infty = \|M_n\|_\infty = 1$.

From Theorem 12.11 we conclude that

$$(I - \tilde{A}_n) = R_n(I - A_n)M_n$$

and

$$(I - \tilde{A}_n)^{-1} = R_n(I - A_n)^{-1}M_n.$$

From these relations we immediately obtain the following theorem.

Theorem 12.23 *For the Nyström method the condition numbers for the linear system are uniformly bounded.*

This theorem states that the Nyström method essentially preserves the stability of the original integral equation.

For the collocation method, we introduce the matrices E_n with entries $u_k(x_j)$ and $\tilde{A}_n$ with entries $(Au_k)(x_j)$. Since $X_n = \operatorname{span}\{u_0, \dots, u_n\}$ is assumed to be such that the interpolation problem with respect to the collocation points $x_0, \dots, x_n$ is uniquely solvable, the matrix E_n is invertible (see Problem 8.1). In addition, let the operator $W_n : \mathbb{R}^{n+1} \to C[a, b]$ be defined by

$$W_n : \gamma \mapsto \sum_{k=0}^{n} \gamma_k u_k$$

for $\gamma = (\gamma_0, \dots, \gamma_n)^T$ and recall the operators R_n and M_n from above. Then we have

$$W_n = L_n M_n E_n.$$

From (12.25) we can conclude that

$$(E_n - \tilde{A}_n) = R_n L_n (I - A) W_n$$

and

$$(E_n - \tilde{A}_n)^{-1} = E_n^{-1} R_n (I - L_n A)^{-1} L_n M_n.$$

From these three relations, and the fact that by Theorems 12.7 and 12.16 the sequence of operators $(I - L_n A)^{-1} L_n$ is uniformly bounded, we obtain the following theorem.

Theorem 12.24 *Under the assumptions of Theorem 12.16, for the collocation method the condition number of the linear system satisfies*

$$\operatorname{cond}(E_n - \tilde{A}_n) \le C \|L_n\|_\infty^2 \operatorname{cond} E_n$$

for all sufficiently large n and some constant C.

This theorem suggests that the basis functions must be chosen with caution. For a poor choice, like monomials, the condition number of E_n can grow quite rapidly. However, for the Lagrange basis, i.e., for the linear system (12.26), E_n becomes the identity matrix with condition number one. In addition, $\|L_n\|$ enters in the estimate on the condition number of the linear system, and for example, for polynomial or trigonometric polynomial interpolation we have $\|L_n\| \to \infty$, $n \to \infty$ (see Theorem 8.16).

In the context of stability we will conclude this chapter with a few remarks on integral equations of the first kind.

Theorem 12.25 *Let X and Y be normed spaces and let $A : X \to Y$ be a compact linear operator. Then A has a bounded inverse if and only if X is finite-dimensional.*

Proof. Assume that A has a bounded inverse $A^{-1} : Y \to X$. Then we have $A^{-1}A = I$, and therefore the identity operator must be compact, since the product of a bounded and a compact operator is compact (see Problem 12.2). However, the identity operator on X is compact if and only if X has finite dimension. □

Theorem 12.25 implies that integral equations of the first kind with continuous (or weakly singular) kernels are improperly posed problems in the sense of Hadamard, as described in Chapter 5.

Of course, the ill-posed nature of an equation has consequences for its numerical treatment. The fact that an operator does not have a bounded inverse means that the condition numbers of its finite-dimensional approximations grow with the quality of the approximation. Hence, a careless discretization of ill-posed problems leads to a numerical behavior that at first glance seems to be paradoxical. Namely, increasing the degree of discretization, i.e., increasing the accuracy of the approximation for the operator, will cause the approximate solution to the equation to become less and less reliable. Therefore, straightforward application of the methods described in this chapter to integral equations of the first kind with continuous kernels will generate numerical nonsense.

To make this remark more vivid, we consider the approximate solution of an integral equation of the first kind

$$\int_a^b K(x,y)\varphi(y)\,dy = f(x), \quad x \in [a,b],$$

by the analogue of the linear system (12.14) for the Nyström method, i.e., by

$$\sum_{k=0}^{n} a_k K(x_j, x_k)\varphi_k^{(n)} = f(x_j), \quad j = 0, \dots, n.$$

The equation of the first kind

$$\int_0^1 (x+1)e^{-xy}\varphi(y)dy = 1 - e^{-(x+1)}, \quad 0 \le x \le 1, \tag{12.37}$$

has the unique solution $\varphi(x) = e^{-x}$ (see Problem 12.20). Table 12.6 gives the difference between the exact solution and the solution obtained by the quadrature method using the (composite) trapezoidal rule.

TABLE 12.6. Numerical solution of (12.37)

n	$x=0$	$x=0.5$	$x=1$
4	0.4057	0.3705	0.1704
8	−4.5989	14.6094	−4.4770
16	−8.5957	2.2626	−153.4805
32	3.8965	−32.2907	22.5570
64	−88.6474	−6.4484	−182.6745

We observe that the approximation is completely useless and that in agreement with the above remarks, the quality of the approximation decreases when the accuracy of the quadrature is increased. (Of course, the actual numerical values for the solution of the ill-conditioned linear system of this example will depend on the actual computer and the code for solving the linear system that is used.)

Hence, the numerical solution of integral equations of the first kind with continuous kernels requires regularization methods such as Tikhonov regularization or singular value cutoff, which we discussed in Chapter 5 for the finite-dimensional case. These regularization techniques now, of course, need to be analyzed in an appropriate function space setting. We recall the corresponding references to [14, 22, 28, 37, 39, 43] from Chapter 5 for the foundation of regularization methods in Hilbert spaces.

Problems

12.1 Show that the boundary value problem for the differential equation

$$-u'' + qu = r \quad \text{in } [0,1]$$

with boundary conditions $u(0) = u(1) = 0$ is equivalent to finding a continuous solution of the integral equation of the second kind

$$u(x) + \int_0^1 G(x,y)q(y)u(y)\,dy = \int_0^1 G(x,y)r(y)\,dy, \quad x \in [0,1],$$

where

$$G(x,y) := \begin{cases} (1-x)y, & 0 \le y \le x \le 1, \\ \\ (1-y)x, & 0 \le x \le y \le 1, \end{cases}$$

is the so-called *Green's function* of the boundary value problem.

12.2 Show that linear combinations of compact linear operators are compact and that the product of two bounded linear operators is compact if one of the factors is compact.

12.3 Show that the integral operator with continuous kernel is a compact operator from $L^2[a,b]$ into $L^2[a,b]$.

12.4 Show that the *Volterra integral equation* of the second kind

$$\varphi(x) - \int_a^x K(x,y)\varphi(y)\,dy = f(x), \quad x \in [a,b],$$

with continuous kernel K has a unique continuous solution φ for each continuous right-hand side f.
Hint: Show that the homogeneous equation allows only the trivial solution and use Theorem 12.2.

12.5 Solve the Volterra integral equation

$$\varphi(x) - \int_0^x e^{x-y}\varphi(y)dy = f(x)$$

by successive approximations.

12.6 Show that a sequence $A_n : X \to Y$ of compact linear operators mapping a normed space X into a normed space Y is collectively compact if and only if for each bounded sequence (φ_n) in X the sequence $(A_n\varphi_n)$ contains a convergent subsequence.

12.7 Show that a sequence (φ_n) of functions $\varphi_n : [a,b] \to \mathbb{R}$ that is equicontinuous and converges pointwise on $[a,b]$ to some function $\varphi : [a,b] \to \mathbb{R}$ converges uniformly on $[a,b]$.

12.8 Prove the *Banach–Steinhaus theorem*: Let $A : X \to Y$ be a bounded linear operator and let $A_n : X \to Y$ be a sequence of bounded linear operators from a Banach space X into a normed space Y. For pointwise convergence $A_n\varphi \to A\varphi$, $n \to \infty$, for all $\varphi \in X$ it is necessary and sufficient that $\|A_n\| \le C$ for all $n \in \mathbb{N}$ with some constant C and that $A_n\varphi \to A\varphi$, $n \to \infty$, for all $\varphi \in U$, where U is some dense subset of X (compare Theorem 9.10).

12.9 For the integral operator A and the numerical integration operators using the (composite) trapezoidal rule, derive bounds on $\|(A_n - A)A\|_\infty$ and $\|(A_n - A)A_n\|_\infty$. Relate the results to Lemma 12.9.

12.10 Verify the integrals (12.21).

12.11 Write a computer program for the Nystöm method allowing the use of different quadrature formulae and test it for various examples.

12.12 Use the quadrature formula (9.36) with the substitution (9.47) in a Nyström method for the integral equation (12.19). Compare the numerical results with those obtained from the trapezoidal and Simpson's rule.

12.13 Verify the Lagrange basis (12.35) and the integrals (12.36).

12.14 Show that the Fourier series of a continuously differentiable periodic function is uniformly convergent.

12.15 In the *degenerate kernel approximation* the integral equation of the second kind with continuous kernel K is approximated by the solutions of

$$\varphi_n(x) - \int_a^b K_n(x,y)\varphi_n(y)\,dy = f(x), \quad x \in [a,b],$$

with an approximate kernel K_n of the form

$$K_n(x,y) = \sum_{j=0}^{n} a_j(x)b_j(y).$$

Show how the solution of the approximate equation can be reduced to solving a system of linear equations. Give an error and convergence analysis based on Theorem 12.6.

12.16 Use the results of Problem 12.15 to prove Theorem 12.2 for the case of an integral equation of the second kind with continuous kernel.

12.17 Construct degenerate kernels via interpolation of the kernel K with respect to the first variable and relate this particular degenerate kernel method to the collocation method (see Problem 12.15).

12.18 The idea of two-grid and multigrid iterations can also be applied to integral equations of the second kind. For its theoretical foundation assume the sequence of operators $A_n : X \to X$ to be either norm convergent (i.e., $\|A_n - A\| \to 0$, $n \to \infty$) or collectively compact and pointwise convergent (i.e., $A_n\varphi \to A\varphi$, $n \to \infty$, for all $\varphi \in X$). Show that the defect correction iteration

$$\varphi_{n,\nu+1} := (I - A_{n-1})^{-1}\{(A_n - A_{n-1})\varphi_{n,\nu} + f_n\}, \quad \nu = 0,1,2,\dots,$$

using the preceding coarser level converges, provided that n is sufficiently large. Show that the defect correction iteration

$$\varphi_{n,\nu+1} := (I - A_0)^{-1}\{(A_n - A_0)\varphi_{n,\nu} + f_n\}, \quad \nu = 0,1,2,\dots,$$

using the coarsest level converges, provided that the approximation A_0 is sufficiently close to A.

12.19 Consider the two-grid iteration

$$\varphi_{n,\nu+1} := (I - A_m)^{-1}\{(A_n - A_m)\varphi_{n,\nu} + f_n\}, \quad \nu = 0, 1, 2, \dots,$$

with $m = n - 1$ or $m = 0$ for the Nyström method, i.e., for the numerical quadrature operators

$$(A_n\varphi)(x) = \sum_{k=1}^{z_n} a_k^{(n)} K(x, x_k^{(n)})\varphi(x_k^{(n)}), \quad x \in [0, 1],$$

with z_n quadrature points. Show that each iteration step requires the following computations. First

$$g_{n,\nu} := f_n + (A_n - A_m)\varphi_{n,\nu}$$

has to be evaluated at the z_m quadrature points $x_j^{(m)}$, $j = 1, \dots, z_m$, on the level m and at the z_n quadrature points $x_j^{(n)}$, $j = 1, \dots, z_n$, on the level n by setting $x = x_j^{(m)}$ and $x = x_j^{(n)}$, respectively, in

$$g_{n,\nu}(x) = f_n(x) + \sum_{k=1}^{z_n} a_k^{(n)} K(x, x_k^{(n)})\varphi_{n,\nu}(x_k^{(n)}) - \sum_{k=1}^{z_m} a_k^{(m)} K(x, x_k^{(m)})\varphi_{n,\nu}(x_k^{(m)}).$$

Then one has to solve the linear system

$$\varphi_{n,\nu+1}(x_j^{(m)}) - \sum_{k=1}^{z_m} a_k^{(m)} K(x_j^{(m)}, x_k^{(m)})\varphi_{n,\nu+1}(x_k^{(m)}) = g_{n,\nu}(x_j^{(m)}), \quad j = 1, \dots, z_m,$$

for the values $\varphi_{n,\nu+1}(x_j^{(m)})$ at the z_m quadrature points $x_j^{(n)}$. Finally, the values at the z_n quadrature points $x_j^{(n)}$, $j = 1, \dots, z_n$, are obtained from the Nyström interpolation

$$\varphi_{n,\nu+1}(x_j^{(n)}) = \sum_{k=1}^{z_m} a_k^{(m)} K(x_j^{(n)}, x_k^{(m)})\varphi_{n,\nu+1}(x_k^{(m)}) + g_{n,\nu}(x_j^{(n)}), \quad j = 1, \dots, z_n.$$

Make an operation count for one step of the defect correction iteration. Set up the corresponding equations for the collocation method.

12.20 Show that the integral equation (12.37) has a unique solution.

References

[1] Anderssen, R.S., de Hoog, F.R., and Lukas, M.A. *The Application and Numerical Solution of Integral Equations.* Sijthoff and Noordhoff, Alphen aan den Rijn 1980.

[2] Anselone, P.M. *Collectively Compact Operator Approximation Theory and Applications to Integral Equations.* Prentice–Hall, Englewood Cliffs 1971.

[3] Atkinson, K.E. *A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind.* SIAM, Philadelphia 1976.

[4] Aubin, J.P. *Approximation of Elliptic Boundary Value Problems.* John Wiley & Sons, New York 1972.

[5] Aubin, J.P. *Applied Functional Analysis.* John Wiley & Sons, New York 1979.

[6] Baker, C.T.H. *The Numerical Treatment of Integral Equations.* Clarendon Press, Oxford 1977.

[7] Ben-Israel, A. and Greville, T.N.E. *Generalized Inverses: Theory and Applications.* John Wiley & Sons, New York 1974.

[8] Brandt, A. Multigrid adaptive solutions to boundary value problems, Math. Comp. **31**, 333–390 (1977).

[9] Brass, H. *Quadraturverfahren.* Vandenhoek und Ruprecht, Göttingen 1979.

[10] Brosowski, B. and Kress, R. *Einführung in die Numerische Mathematik.* Bibliographisches Institut, Mannheim 1975.

[11] Ciarlet, P.S. *The Finite Element Method for Elliptic Problems.* North Holland, Amsterdam 1978.

[12] Coddington, E.A. and Levinson, N. *Theory of Ordinary Differential Equations.* McGraw-Hill, New York 1955.

[13] Collatz, L. *The Numerical Treatment of Differential Equations.* 3rd edition. Springer-Verlag, Berlin 1966.

[14] Colton, D. and Kress, R. *Inverse Acoustic and Electromagnetic Scattering Theory.* 2nd edition. Springer-Verlag, Berlin 1998.

[15] Davis, P.J. On the numerical integration of periodic analytic functions. In: *Symposium on Numerical Approximation* (R. Langer, ed.). The University of Wisconsin Press, Madison, 45–59 (1959).

[16] Davis, P.J. *Interpolation and Approximation.* Blaisdell Publishing Company, Waltham 1963.

[17] Davis, P.J. and Rabinowitz, P. *Methods of Numerical Integration.* 2nd edition. Academic Press, San Diego 1984.

[18] De Boor, C. *A Practical Guide to Splines.* Springer-Verlag, New York 1978.

[19] Delves, L.M. and Mohamed, J.L. *Computational Methods for Integral Equations.* Cambridge University Press, Cambridge 1985.

[20] Dennis, J.E. and Schnabel, R.B. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Prentice-Hall, Englewood Cliffs 1983.

[21] Engels, H. *Numerical Quadrature and Cubature.* Academic Press, New York 1980.

[22] Engl, H.W., Hanke, M., and Neubauer, A. *Regularization of Inverse Problems.* Kluwer Academic Publishers, Dordrecht 1996.

[23] Farin, G. *Curves and Surfaces for Computer Aided Geometric Design. A Practical Guide.* 2nd edition. Academic Press, Boston 1990.

[24] Gilbarg, D. and Trudinger, N.S. *Elliptic Partial Differential Equations of Second Order.* Springer-Verlag, Berlin 1977.

[25] Golberg, M.A. and Chen, C.S. *Discrete Projection Methods for Integral Equations.* Computational Mechanics Publications, Southampton 1997.

[26] Golub, G. and Ortega, J.M. *Scientific Computing.* Academic Press, Boston 1993.

[27] Golub, G. and van Loan, C. *Matrix Computations.* John Hopkins University Press, Baltimore 1989.

[28] Groetsch, C.W. *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind.* Pitman, Boston 1984.

[29] Hackbusch, W. *Multi-Grid Methods and Applications.* Springer-Verlag, Berlin 1985.

[30] Hackbusch, W. *Integral Equations: Theory and Numerical Treatment.* Birkhäuser-Verlag, Basel 1995.

[31] Hadamard, J. *Lectures on Cauchy's Problem in Linear Partial Differential Equations.* Yale University Press, New Haven 1923.

[32] Hairer, E., Nørsett, S.P., and Wanner, G. *Solving Ordinary Differential Equations. Nonstiff Problems.* Springer-Verlag, Berlin 1987.

[33] Henrici, P. *Discrete Variable Methods in Ordinary Differential Equations.* John Wiley & Sons, New York 1962.

[34] Heuser, H. *Funktionalanalysis.* 2. Auflage. Teubner, Stuttgart 1986.

[35] Kantorovic, L.V. and Akilov, G.P. *Functional Analysis in Normed Spaces.* Pergamon Press, Oxford 1964.

[36] Keller, H.B. *Numerical Methods for Two-Point Boundary Value Problems.* Blaisdell Publishing Company, Waltham 1968.

[37] Kirsch, A. *An Introduction to the Mathematical Theory of Inverse Problems.* Springer-Verlag, New York 1996.

[38] Kress, R. Ein ableitungsfreies Restglied für die trigonometrische Interpolation periodischer analytischer Funktionen. Numer. Math. **16**, 389–396 (1971).

[39] Kress, R. *Linear Integral Equations.* Springer-Verlag, Berlin 1989.

[40] Kress, R. A Nyström method for boundary integral equations in domains with corners. Numer. Math. **58**, 145–161 (1990).

[41] Kress, R., de Vries, H.L., and Wegmann, R. On nonnormal matrices. Linear Algebra and its Appl. **8**, 109–120 (1974).

[42] Lambert, J.D. *Numerical Methods for Ordinary Differential Equations. The Initial Value Problem.* John Wiley & Sons, Chichester 1993.

[43] Louis, A.K. *Inverse und schlecht gestellte Probleme.* Teubner, Stuttgart 1989.

[44] Moré, J.J. The Levenberg–Marquardt algorithm, implementation and theory. In: *Numerical Analysis* (Watson, ed.). Springer-Verlag Lecture Notes in Mathematics **630**, Berlin, 105–116 (1977).

[45] Nussbaumer, H.J. *Fast Fourier Transform and Convolution Algorithms.* Springer-Verlag, Berlin 1982.

[46] Ortega, J.M. and Poole, W.G. *An Introduction to Numerical Methods for Differential Equations.* Pitman, Boston 1981.

[47] Ortega, J.M. and Rheinboldt, W.C. *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, New York 1970.

[48] Parlett, B.N. *The Symmetric Eigenvalue Problem.* Prentice-Hall, Englewood Cliffs 1980.

[49] Prössdorf, S. and Silbermann, B. *Numerical Analysis for Integral and Related Operator Equations.* Akademie-Verlag, Berlin 1991, and Birkhäuser-Verlag, Basel 1991.

[50] Roberts, S.M. and Shipman, J.S. *Two-Point Boundary Value Problems: Shooting Methods.* Elsevier, New York 1972.

[51] Rudin, W. *Functional Analysis.* McGraw-Hill, New York 1973.

[52] Sag, T.W. and Szegeres, G. Numerical evaluation of high-dimensional integrals. Math. Comp. **18**, 245–253 (1964).

[53] Schumaker, L.L. *Spline Functions: Basic Theory.* John Wiley & Sons, Chichester 1981.

[54] Sidi, A. A new variable transformation for numerical integration. In: *Numerical Integration IV.* (Brass, Hämmerlin, eds.) International Series of Numerical Mathematics. Birkhäuser-Verlag Basel **112**, 359–373 (1993).

[55] Stetter, H.J. *Analysis of Discretization Methods for Ordinary Differential Equations.* Springer-Verlag, Berlin 1973.

[56] Stetter, H. J. The defect correction principle and discretization methods, Numer. Math. **29**, 425–443 (1978).

[57] Stroud, A.H. *Approximate Calculation of Multiple Integrals.* Prentice-Hall, Englewood Cliffs 1971.

[58] Takahasi, H. and Mori, M. Quadrature formulas obtained by variable transformation. Numer. Math. **21**, 206–219 (1973).

[59] Taylor, A.E. *Introduction to Functional Analysis.* John Wiley & Sons, New York 1967.

[60] Treves, F. *Basic Linear Partial Differential Equations.* Academic Press, New York 1975.

[61] Varga, R. *Matrix Iterative Analysis.* Prentice-Hall, Englewood Cliffs 1962.

[62] Watkins, D.S. Understanding the QR-Algorithm, SIAM Review **24**, (1982).

[63] Weissinger, J. *Spärlich besetzte Gleichungssysteme.* Bibliographisches Institut, Mannheim 1990.

[64] Wesseling, P. *An Introduction to Multigrid Methods.* John Wiley & Sons, Chichester 1992.

[65] Wilkinson, J.H. *The Algebraic Eigenvalue Problem.* Clarendon Press, Oxford 1965.

[66] Young, D. *Iterative Solution of Large Linear Systems.* Academic Press, New York 1971.

Index

Graduate Texts in Mathematics

continued from page ii

119 ROTMAN. An Introduction to Algebraic Topology.
120 ZIEMER. Weakly Differentiable Functions: Sobolev Spaces and Functions of Bounded Variation.
121 LANG. Cyclotomic Fields I and II. Combined 2nd ed.
122 REMMERT. Theory of Complex Functions.
Readings in Mathematics
123 EBBINGHAUS/HERMES et al. Numbers.
Readings in Mathematics
124 DUBROVIN/FOMENKO/NOVIKOV. Modern Geometry—Methods and Applications. Part III.
125 BERENSTEIN/GAY. Complex Variables: An Introduction.
126 BOREL. Linear Algebraic Groups. 2nd ed.
127 MASSEY. A Basic Course in Algebraic Topology.
128 RAUCH. Partial Differential Equations.
129 FULTON/HARRIS. Representation Theory: A First Course.
Readings in Mathematics
130 DODSON/POSTON. Tensor Geometry.
131 LAM. A First Course in Noncommutative Rings.
132 BEARDON. Iteration of Rational Functions.
133 HARRIS. Algebraic Geometry: A First Course.
134 ROMAN. Coding and Information Theory.
135 ROMAN. Advanced Linear Algebra.
136 ADKINS/WEINTRAUB. Algebra: An Approach via Module Theory.
137 AXLER/BOURDON/RAMEY. Harmonic Function Theory.
138 COHEN. A Course in Computational Algebraic Number Theory.
139 BREDON. Topology and Geometry.
140 AUBIN. Optima and Equilibria. An Introduction to Nonlinear Analysis.
141 BECKER/WEISPFENNING/KREDEL. Gröbner Bases. A Computational Approach to Commutative Algebra.
142 LANG. Real and Functional Analysis. 3rd ed.
143 DOOB. Measure Theory.
144 DENNIS/FARB. Noncommutative Algebra.
145 VICK. Homology Theory. An Introduction to Algebraic Topology. 2nd ed.
146 BRIDGES. Computability: A Mathematical Sketchbook.
147 ROSENBERG. Algebraic *K*-Theory and Its Applications.
148 ROTMAN. An Introduction to the Theory of Groups. 4th ed.
149 RATCLIFFE. Foundations of Hyperbolic Manifolds.
150 EISENBUD. Commutative Algebra with a View Toward Algebraic Geometry.
151 SILVERMAN. Advanced Topics in the Arithmetic of Elliptic Curves.
152 ZIEGLER. Lectures on Polytopes.
153 FULTON. Algebraic Topology: A First Course.
154 BROWN/PEARCY. An Introduction to Analysis.
155 KASSEL. Quantum Groups.
156 KECHRIS. Classical Descriptive Set Theory.
157 MALLIAVIN. Integration and Probability.
158 ROMAN. Field Theory.
159 CONWAY. Functions of One Complex Variable II.
160 LANG. Differential and Riemannian Manifolds.
161 BORWEIN/ERDÉLYI. Polynomials and Polynomial Inequalities.
162 ALPERIN/BELL. Groups and Representations.
163 DIXON/MORTIMER. Permutation Groups.
164 NATHANSON. Additive Number Theory: The Classical Bases.
165 NATHANSON. Additive Number Theory: Inverse Problems and the Geometry of Sumsets.
166 SHARPE. Differential Geometry: Cartan's Generalization of Klein's Erlangen Program.
167 MORANDI. Field and Galois Theory.
168 EWALD. Combinatorial Convexity and Algebraic Geometry.
169 BHATIA. Matrix Analysis.
170 BREDON. Sheaf Theory. 2nd ed.
171 PETERSEN. Riemannian Geometry.
172 REMMERT. Classical Topics in Complex Function Theory.
173 DIESTEL. Graph Theory.
174 BRIDGES. Foundations of Real and Abstract Analysis.
175 LICKORISH. An Introduction to Knot Theory.
176 LEE. Riemannian Manifolds.
177 NEWMAN. Analytic Number Theory.
178 CLARKE/LEDYAEV/STERN/WOLENSKI. Nonsmooth Analysis and Control Theory.
180 SRIVASTAVA. A Course on Borel Sets.
181 KRESS. Numerical Analysis.